International Journal of Artificial Intelligence in Education (2020) 30:121–204
https://doi.org/10.1007/s40593-019-00186-y

ARTICLE

A Systematic Review of Automatic Question Generation for Educational Purposes

Ghader Kurdi[1] · Jared Leo[1] · Bijan Parsia[1] · Uli Sattler[1] · Salam Al-Emari[2]

Published online: 21 November 2019

Abstract

While exam-style questions are a fundamental educational tool serving a variety of purposes, manual construction of questions is a complex process that requires training, experience, and resources. This, in turn, hinders and slows down the use of educational activities (e.g. providing practice questions) and new advances (e.g. adaptive testing) that require a large pool of questions. To reduce the expenses associated with manual construction of questions and to satisfy the need for a continuous supply of new questions, automatic question generation (AQG) techniques were introduced. This review extends a previous review on AQG literature that has been published up to late 2014. It includes 93 papers that were between 2015 and early 2019 and tackle the automatic generation of questions for educational purposes. The aims of this review are to: provide an overview of the AQG community and its activities, summarise the current trends and advances in AQG, highlight the changes that the area has undergone in the recent years, and suggest areas for improvement and future opportunities for AQG. Similar to what was found previously, there is little focus in the current literature on generating questions of controlled difficulty, enriching question forms and structures, automating template construction, improving presentation, and generating feedback. Our findings also suggest the need to further improve experimental reporting, harmonise evaluation metrics, and investigate other evaluation methods that are more feasible.

Keywords Automatic question generation · Semantic Web · Education · Natural language processing · Natural language generation · Assessment · Difficulty prediction

✉ Ghader Kurdi
ghader.kurdi@manchester.ac.uk

Extended author information available on the last page of the article.

Introduction

Exam-style questions are a fundamental educational tool serving a variety of purposes. In addition to their role as an assessment instrument, questions have the potential to influence student learning. According to Thalheimer (2003), some of the benefits of using questions are: 1) offering the opportunity to practice retrieving information from memory; 2) providing learners with feedback about their misconceptions; 3) focusing learners' attention on the important learning material; 4) reinforcing learning by repeating core concepts; and 5) motivating learners to engage in learning activities (e.g. reading and discussing). Despite these benefits, manual question construction is a challenging task that requires training, experience, and resources. Several published analyses of real exam questions (mostly multiple choice questions (MCQs)) (Hansen and Dexter 1997; Tarrant et al. 2006; Hingorjo and Jaleel 2012; Rush et al. 2016) demonstrate their poor quality, which Tarrant et al. (2006) attributed to a lack of training in assessment development. This challenge is augmented further by the need to replace assessment questions consistently to ensure their validity, since their value will decrease or be lost after a few rounds of usage (due to being shared between test takers), as well as the rise of e-learning technologies, such as massive open online courses (MOOCs) and adaptive learning, which require a larger pool of questions.

Automatic question generation (AQG) techniques emerged as a solution to the challenges facing test developers in constructing a large number of good quality questions. AQG is concerned with the construction of algorithms for producing questions from knowledge sources, which can be either structured (e.g. knowledge bases (KBs) or unstructured (e.g. text)). As Alsubait (2015) discussed, research on AQG goes back to the 70's. Nowadays, AQG is gaining further importance with the rise of MOOCs and other e-learning technologies (Qayyum and Zawacki-Richter 2018; Gaebel et al. 2014; Goldbach and Hamza-Lup 2017).

In what follows, we outline some potential benefits that one might expect from successful automatic generation of questions. AQG can reduce the cost (in terms of both money and effort) of question construction which, in turn, enables educators to spend more time on other important instructional activities. In addition to resource saving, having a large number of good-quality questions enables the enrichment of the teaching process with additional activities such as adaptive testing (Vie et al. 2017), which aims to adapt learning to student knowledge and needs, as well as drill and practice exercises (Lim et al. 2012). Finally, being able to automatically control question characteristics, such as question difficulty and cognitive level, can inform the construction of good quality tests with particular requirements.

Although the focus of this review is education, the applications of question generation (QG) are not limited to education and assessment. Questions are also generated for other purposes, such as validation of knowledge bases, development of conversational agents, and development of question answering or machine reading comprehension systems, where questions are used for training and testing.

This review extends a previous systematic review on AQG (Alsubait 2015), which covers the literature up to the end of 2014. Given the large amount of research that has been published since Alsubait's review was conducted (93 papers over a four year period compared to 81 papers over the preceding 45-year period), an extension of Alsubait's review is reasonable at this stage. To capture the recent developments in the field, we review the literature on AQG from 2015 to early 2019. We take Alsubait's review as a starting point and extend the methodology in a number of ways (e.g. additional review questions and exclusion criteria), as will be described in the sections titled "Review Objective" and "Review Method". The contribution of this review is in providing researchers interested in the field with the following:

1. a comprehensive summary of the recent AQG approaches;
2. an analysis of the state of the field focusing on differences between the pre- and post-2014 periods;
3. a summary of challenges and future directions; and
4. an extensive reference to the relevant literature.

Summary of Previous Reviews

There have been six published reviews on the AQG literature. The reviews reported by Le et al. 2014, Kaur and Bathla 2015, Alsubait 2015 and Rakangor and Ghodasara (2015) cover the literature that has been published up to late 2014 while those reported by Ch and Saha (2018) and Papasalouros and Chatzigiannakou (2018) cover the literature that has been published up to late 2018. Out of these, the most comprehensive review is Alsubait's, which includes 81 papers (65 distinct studies) that were identified using a systematic procedure. The other reviews were selective and only cover a small subset of the AQG literature. Of interest, due to it being a systematic review and due to the overlap in timing with our review, is the review developed by Ch and Saha (2018). However, their review is not as rigorous as ours, as theirs only focuses on automatic generation of MCQs using text as input. In addition, essential details about the review procedure, such as the search queries used for each electronic database and the resultant number of papers, are not reported. In addition, several related studies found in other reviews on AQG are not included.

Findings of Alsubait's Review

In this section, we concentrate on summarising the main results of Alsubait's systematic review, due to its being the only comprehensive review. We do so by elaborating on interesting trends and speculating about the reasons for those trends, as well as highlighting limitations observed in the AQG literature.

Alsubait characterised AQG studies along the following dimensions: 1) purpose of generating questions, 2) domain, 3) knowledge sources, 4) generation method, 5) question type, 6) response format, and 7) evaluation.

The results of the review and the most prevalent categories within each dimension are summarised in Table 1. As can be seen in Table 1, generating questions for a specific domain is more prevalent than generating domain-unspecific questions. The most investigated domain is language learning (20 studies), followed by mathematics and medicine (four studies each). Note that, for these three domains,

Table 1 Results of Alsubait's review. Categories with frequency of three or less are classified under "other"

Dimension	Categories	No. of studies	Percentage
Purpose	Assessment	51	78.5%
	Knowledge acquisition	7	10.8%
	Validation	4	6.2%
	General	3	4.6%
Domain	Domain-specific	35	53.9%
	Generic	30	46.2%
Knowledge source	Text	38	58.5%
	Ontologies	11	16.9%
	Other	16	24.6%
Generation Method	Syntax based	26	38.2%
	Semantic based	25	36.8%
	Template based	12	17.7%
	Other	5	7.4%
Question type	Factual wh-questions	21	30.0%
	Fill-in-the-blank questions	17	24.3%
	Math word problems	4	5.7%
	Other	28	40.0%
Response format	Free response	33	50.8%
	Multiple choice	31	47.7%
	True/false	1	1.5%
Evaluation	Expert-centred	20	30.8%
	Student-centred	15	23.1%
	Other	12	18.5%
	None	18	27.7%

there are large standardised tests developed by professional organisations (e.g. Test of English as a Foreign Language (TOEFL), International English Language Testing System (IELTS) and Test of English for International Communication (TOEIC) for language, Scholastic Aptitude Test (SAT) for mathematics and board examinations for medicine). These tests require a continuous supply of new questions. We believe that this is one reason for the interest in generating questions for these domains. We also attribute the interest in the language learning domain to the ease of generating language questions, relative to questions belonging to other domains. Generating language questions is easier than generating other types of questions for two reasons: 1) the ease of adopting text from a variety of publicly available resources (e.g. a large number of general or specialised textual resources can be used for reading comprehension (RC)) and 2) the availability of natural language processing (NLP) tools for shallow understanding of text (e.g. part of speech (POS) tagging) with an acceptable performance, which is often sufficient for generating language questions.

To illustrate, in Chen et al. (2006), the distractors accompanying grammar questions are generated by changing the verb form of the key (e.g. “write”, “written”, and “wrote” are distractors while “writing” is the key). Another plausible reason for interest in questions on medicine is the availability of NLP tools (e.g. named entity recognisers and co-reference resolvers) for processing medical text. There are also publicly available knowledge bases, such as UMLS (Bodenreider 2004) and SNOMED-CT (Donnelly 2006), that are utilised in different tasks such as text annotation and distractor generation. The other investigated domains are analytical reasoning, geometry, history, logic, programming, relational databases, and science (one study each).

With regard to knowledge sources, the most commonly used source for question generation is text (Table 1). A similar trend was also found by Rakangor and Ghodasara (2015). Note that 19 text-based approaches, out of the 38 text-based approaches identified by Alsubait (2015), tackle the generation of questions for the language learning domain, both free response (FR) and multiple choice (MC). Out of the remaining 19 studies, only five focus on generating MCQs. To do so, they incorporate additional inputs such as WordNet (Miller et al. 1990), thesaurus, or textual corpora. By and large, the challenge in the case of MCQs is distractor generation. Despite using text for generating language questions, where distractors can be generated using simple strategies such as selecting words having a particular POS or other syntactic properties, text often does not incorporate distractors, so external, structured knowledge sources are needed to find what is true and what is similar. On the other hand, eight ontology-based approaches are centred on generating MCQs and only three focus on FR questions.

Simple factual wh-questions (i.e. where the answers are short facts that are explicitly mentioned in the input) and gap-fill questions (also known as fill-in-the-blank or cloze questions) are the most generated types of questions with the majority of them, 17 and 15 respectively, being generated from text. The prevalence of these questions is expected because they are common in language learning assessment. In addition, these two types require relatively little effort to construct, especially when they are not accompanied by distractors. In gap-fill questions, there are no concerns about the linguistic aspects (e.g. grammaticality) because the stem is constructed by only removing a word or a phrase from a segment of text. The stem of a wh-question is constructed by removing the answer from the sentence, selecting an appropriate wh-word, and rearranging words to form a question. Other types of questions such as mathematical word problems, Jeopardy-style questions,[1] and medical case-based questions (CBQs) require more effort in choosing the stem content and verbalisation. Another related observation we made is that the types of questions generated from ontologies are more varied than the types of questions generated from text.

Limitations observed by Alsubait (2015) include the limited research on controlling the difficulty of generated questions and on generating informative feedback.

[1] Questions like those presented in the T.V. show “Jeopardy!”. These questions consist of statements that give hints about the answer. See Faizan and Lohmann (2018) for an example.

Existing difficulty models are either not validated or only applicable to a specific type of question (Alsubait 2015). Regarding feedback (i.e. an explanation for the correctness/incorrectness of the answer), only three studies generate feedback along with the questions. Even then, the feedback is used to motivate students to try again or to provide extra reading material without explaining why the selected answer is correct/incorrect. Ungrammaticality is another notable problem with auto-generated questions, especially in approaches that apply syntactic transformations of sentences (Alsubait 2015). For example, 36.7% and 39.5% of questions generated in the work of Heilman and Smith (2009) were rated by reviewers as ungrammatical and nonsensical, respectively. Another limitation related to approaches to generating questions from ontologies is the use of experimental ontologies for evaluation, neglecting the value of using existing, probably large, ontologies. Various issues can arise if existing ontologies are used, which in turn provide further opportunities to enhance the quality of generated questions and the ontologies used for generation.

Review Objective

The goal of this review is to provide a comprehensive view of the AQG field since 2015. Following and extending the schema presented by Alsubait (2015) (Table 1), we have structured our review around the following four objectives and their related questions. Questions marked with an asterisk "*" are those proposed by Alsubait (2015). Questions under the first three objectives (except question 5 under OBJ3) are used to guide data extraction. The others are analytical questions to be answered based on extracted results.

OBJ1: Providing an overview of the AQG community and its activities

1. What is the rate of publication?*
2. What types of papers are published in the area?
3. Where is research published?
4. Who are the active research groups in the field?*

OBJ2: Summarising current QG approaches

1. What is the purpose of QG?*
2. What method is applied?*
3. What tasks related to question generation are considered?
4. What type of input is used?*
5. Is it designed for a specific domain? For which domain?*
6. What type of questions are generated?* (i.e., question format and answer format)
7. What is the language of the questions?
8. Does it generate feedback?*
9. Is difficulty of questions controlled?*
10. Does it consider verbalisation (i.e. presentation improvements)?

OBJ3: Identifying the gold-standard performance in AQG

1. Are there any available sources or standard datasets for performance comparison?
2. What types of evaluation are applied to QG approaches?*
3. What properties of questions are evaluated?[2] and What metrics are used for their measurement?
4. How does the generation approach perform?
5. What is the gold-standard performance?

OBJ4: Tracking the evolution of AQG since Alsubait's review

1. Has there been any progress on feedback generation?
2. Has there been progress on generating questions with controlled difficulty?
3. Has there been progress on enhancing the naturalness of questions (i.e. verbalisation)?

One of our motivations for pursuing these objectives is to provide members of the AQG community with a reference to facilitate decisions such as what resources to use, whom to compare to, and where to publish. As we mentioned in the Summary of Previous Reviews, Alsubait (2015) highlighted a number of concerns related to the quality of generated questions, difficulty models, and the evaluation of questions. We were motivated to know whether these concerns have been addressed. Furthermore, while reviewing some of the AQG literature, we made some observations about the simplicity of generated questions and about the reporting being insufficient and heterogeneous. We want to know whether these issues are universal across the AQG literature.

Review Method

We followed the systematic review procedure explained in (Kitchenham and Charters 2007; Boland et al. 2013).

Inclusion and Exclusion Criteria

We included studies that tackle the generation of questions for educational purposes (e.g. tutoring systems, assessment, and self-assessment) without any restriction on domains or question types. We adopted the exclusion criteria used in Alsubait (2015) (1 to 5) and added additional exclusion criteria (6 to 13). A paper is excluded if:

1. it is not in English
2. it presents work in progress only and does not provide a sufficient description of how the questions are generated

[2] Note that evaluated properties are not necessarily controlled by the generation method. For example, an evaluation could focus on difficulty and discrimination as an indication of quality.

3. it presents a QG approach that is based mainly on a template and questions are generated by substituting template slots with numerals or with a set of randomly predefined values
4. it focuses on question answering rather than question generation
5. it presents an automatic mechanism to deliver assessments, rather than generating assessment questions
6. it presents an automatic mechanism to assemble exams or to adaptively select questions from a question bank
7. it presents an approach for predicting the difficulty of human-authored questions
8. it presents a QG approach for purposes other than those related to education (e.g. training of question answering systems, dialogue systems)
9. it does not include an evaluation of the generated questions
10. it is an extension of a paper published before 2015 and no changes were made to the question generation approach
11. it is a secondary study (i.e. literature review)
12. it is not peer-reviewed (e.g. theses, presentations and technical reports)
13. its full text is not available (through the University of Manchester Library website, Google or Google scholar).

Search Strategy

Data Sources Six data sources were used, five of which were electronic databases (ERIC, ACM, IEEE, INSPEC and Science Direct), which were determined by Alsubait (2015) to have good coverage of the AQG literature. We also searched the International Journal of Artificial Intelligence in Education (AIED) and the proceedings of the International Conference on Artificial Intelligence in Education for 2015, 2017, and 2018 due to their AQG publication record.

We obtained additional papers by examining the reference lists of, and the citations to, AQG papers we reviewed (known as “snowballing”). The citations to a paper were identified by searching for the paper using Google Scholar, then clicking on the “cited by” option that appears under the name of the paper. We performed this for every paper on AQG, regardless of whether we had decided to include it, to ensure that we captured all the relevant papers. That is to say, even if a paper was excluded because it met some of the exclusion criteria (1-3 and 8-13), it is still possible that it refers to, or is referred to by, relevant papers.

We used the reviews reported by Ch and Saha (2018) and Papasalouros and Chatzigiannakou (2018) as a “sanity check” to evaluate the comprehensiveness of our search strategy. We exported all the literature published between 2015 and 2018 included in the work of Ch and Saha (2018) and Papasalouros and Chatzigiannakou (2018) and checked whether they were included in our results (both search results and snowballing results).

Search Queries We used the keywords “question” and “generation” to search for relevant papers. Actual search queries used for each of the databases are provided in the Appendix under “Search Queries”. We decided on these queries after experimenting with different combinations of keywords and operators provided by each database and looking at the ratio between relevant and irrelevant results in the first few pages (sorted by relevance). To ensure that recall was not compromised, we checked whether relevant results returned using different versions of each search query were still captured by the selected version.

Screening The search results were exported to comma-separated values (CSV) files. Two reviewers then looked independently at the titles and abstracts to decide on inclusion or exclusion. The reviewers skimmed the paper if they were not able to make a decision based on the title and abstract. Note that, at this phase, it was not possible to assess whether all papers had satisfied the exclusion criteria 2, 3, 8, 9, and 10. Because of this, the final decision was made after reading the full text as described next.

To judge whether a paper's purpose was related to education, we considered the title, abstract, introduction, and conclusion sections. Papers that mentioned many potential purposes for generating questions, but did not state which one was the focus, were excluded. If the paper mentioned only educational applications of QG, we assumed that its purpose was related to education, even without a clear purpose statement. Similarly, if the paper mentioned only one application, we assumed that was its focus.

Concerning evaluation, papers that evaluated the usability of a system that had a QG functionality, without evaluating the quality of generated questions, were excluded. In addition, in cases where we found multiple papers by the same author(s) reporting the same generation approach, even if some did not cover evaluation, all of the papers were included but counted as one study in our analyses.

Lastly, because the final decision on inclusion/exclusion sometimes changed after reading the full paper, agreement between the two reviewers was checked after the full paper had been read and the final decision had been made. However, a check was also made to ensure that the inclusion/exclusion criteria were interpreted in the same way. Cases of disagreement were resolved through discussion.

Data Extraction

Guided by the questions presented in the "Review Objective" section, we designed a specific data extraction form. Two reviewers independently extracted data related to the included studies. As mentioned above, different papers that related to the same study were represented as one entry. Agreement for data extraction was checked and cases of disagreement were discussed to reach a consensus.

Papers that had at least one shared author were grouped together if one of the following criteria were met:

- they reported on different evaluations of the same generation approach;
- they reported on applying the same generation approach to different sources or domains;

Table 2 Criteria used for quality assessment

Participants
Q1: Is the number of the participants included in the study reported?
Q2: Are the characteristics of the participants included in the study described?
Q3: Is the procedure for participant selection reported?
Q4: Are the participants selected for this study suitable for the question(s) posed by the researchers?
Question sample
Q5: Is the number of questions evaluated in the study reported?
Q6: Is the sample selection method described?
Q6a: Is the sampling strategy described?
Q6b: Is the sample size calculation described?
Q7: Is the sample representative of the target group?
Measures used
Q8: Are the main outcomes to be measured described?
Q9: Is the reliability of the measures assessed?

- one of the papers introduced an additional feature of the generation approach such as difficulty prediction or generating distractors without changing the initial generation procedure.

The extracted data were analysed using a code written in R markdown.[3]

Quality Assessment

Since one of the main objectives of this review is to identify the gold standard performance, we were interested in the quality of the evaluation approaches. To assess this, we used the criteria presented in Table 2 which were selected from existing checklists (Downs and Black 1998; Reisch et al. 1989; Critical Appraisal Skills Programme 2018), with some criteria being adapted to fit specific aspects of research on AQG. The quality assessment was conducted after reading a paper and filling in the data extraction form.

In what follows, we describe the individual criteria (Q1-Q9 presented in Table 2) that we considered when deciding if a study satisfied said criteria. Three responses are used when scoring the criteria: “yes”, “no” and “not specified”. The “not specified” response is used when either there is no information present to support the criteria, or when there is not enough information present to distinguish between a “yes” or “no” response.

Q1-Q4 are concerned with the quality of reporting on participant information, Q5-Q7 are concerned with the quality of reporting on the question samples, and Q8 and Q9 describe the evaluative measures used to assess the outcomes of the studies.

[3] The code and the input files are available at: https://github.com/grkurdi/AQG_systematic_review

Q1: When a study reports the exact number of participants (e.g. experts, students, employees, etc.) used in the study, Q1 scores a "yes". Otherwise, it scores a "no". For example, the passage *"20 students were recruited to participate in an exam* ..." would result in a "yes", whereas *"a group of students were recruited to participate in an exam* ..." would result in a "no".

Q2: Q2 requires the reporting of demographic characteristics supporting the suitability of the participants for the task. Depending on the category of participant, relevant demographic information is required to score a "yes". Studies that do not specify relevant information score a "no". By means of examples, in studies relying on expert reviews, those that include information on teaching experience or the proficiency level of reviewers would receive a "yes", while in studies relying on mock exams, those that include information about grade level or proficiency level of test takers would also receive a "yes". Studies reporting that the evaluation was conducted by reviewers, instructors, students, or co-workers without providing any additional information about the suitability of the participants for the task would be considered neglectful of Q2 and score a "no".

Q3: For a study to score "yes" for Q3, it must provide specific information on how participants were selected/recruited, otherwise it receives a score of "no". This includes information on whether the participants were paid for their work or were volunteers. For example, the passage *"7th grade biology students were recruited from a local school."* would receive a score of "no" because it is not clear whether or not they were paid for their work. However, a study that reports *"Student volunteers were recruited from a local school* ..." or *"Employees from company X were employed for n hours to take part in our study... they were rewarded for their services with Amazon vouchers worth $n"* would receive a "yes".

Q4: To score "yes" for Q4, two conditions must be met: the study must 1) score "yes" for both Q2 and Q3 and 2) only use participants that are suitable for the task at hand. Studies that fail to meet the first condition score "not specified" while those that fail to meet the second condition score "no". Regarding the suitability of participants, we consider, as an example, native Chinese speakers suitable for evaluating the correctness and plausibility of options generated for Chinese gap-fill questions. As another example, we consider Amazon Mechanical Turk (AMT) co-workers unsuitable for evaluating the difficulty of domain-specific questions (e.g. mathematical questions).

Q5: When a study reports the exact number of questions used in the experimentation or evaluation stage, Q5 receives a score of "yes", otherwise it receives a score of "no". To demonstrate, consider the following examples. A study reporting *"25 of the 100 generated questions were used in our evaluation...*" would receive a score of "yes". However, if a study made a claim such as *"Around half of the generated questions were used...*", it would receive a score of "no".

Q6: Q6a requires that the sampling strategy be not only reported (e.g. random, proportionate stratification, disproportionate stratification, etc.) but also justified to receive a "yes", otherwise, it receives a score of "no". To demonstrate, if a study only reports that *"We sampled 20 questions from each template ... "* would receive a score of "no" since no justification as to why the stratified sampling procedure was used is provided. However, if it was to also add *"We sampled 20 questions* *from each template to ensure template balance in discussions about the quality of generated questions...*" then this would be considered as a suitable justification and would warrant a score of "yes". Similarly, Q6b requires that the sample size be both reported and justified.

Q7: Our decision regarding Q7 takes into account the following: 1) responses to Q6a (i.e. a study can only score "yes" if the score to Q6a is "yes", otherwise, the score would be "not specified") and 2) representativeness of the population. Using random sampling is, in most cases, sufficient to score "yes" for Q7. However, if multiple types of questions are generated (e.g. different templates or different difficulty levels), stratified sampling is more appropriate in cases in which the distribution of questions is skewed.

Q8: Q8 considers whether the authors provide a description, a definition, or a mathematical formula for the evaluation measures they used as well as a description of the coding system (if applicable). If so, then the study receives a score of "yes" for Q8, otherwise it receives a score of "no".

Q9: Q9 is concerned with whether questions were evaluated by multiple reviewers and whether measures of the agreement (e.g., Cohen's kappa or percentage of agreement) were reported. For example, studies reporting information similar to *"all questions were double-rated and inter-rater agreement was computed... "* receive a score of "yes", whereas studies reporting information similar to *"Each question was rated by one reviewer... "* receive a score of "no" .

To assess inter-rater reliability, this activity was performed by two reviewers (the first and second authors), who are proficient in the field of AQG, independently on an exploratory random sample of 27 studies.[4] The percentage of agreement and Cohen's kappa were used to measure inter-rater reliability for Q1-Q9. The percentage of agreement ranged from 73% to 100%, while Cohen's kappa was above .72 for Q1-Q5, demonstrating "substantial to almost perfect agreement", and equal to 0.42 for Q9,[5]

Results and Discussion

Search and Screening Results

Searching the databases and AIED resulted in 2,012 papers and we checked 974.[7] The difference is due to ACM which provided 1,265 results and we only checked the first 200 results (sorted by relevance) because we found that subsequent results became irrelevant. Out of the search results, 122 papers were considered relevant after

[4] The required sample size was calculated using the N.cohen.kappa function (Gamer et al. 2019).

[5] This due to the initial description of Q9 being insufficient. However, the agreement improved after refining the description of Q9. demonstrating "moderate agreement".[6] Note that Cohen's kappa was unsuitable for assessing the agreement on the criteria Q6-Q8 due to the unbalanced distribution of responses (e.g. the majority of responses to Q6a were "no"). Since the level of agreement between both reviewers was high, the quality of the remaining studies was assessed by the first author.

[7] The last update of the search was on 3-4-2019.

looking at their titles and abstracts. After removing duplicates, 89 papers remained. This set was further reduced to 36 papers after reading the full text of the papers. Checking related work sections and the reference lists identified 169 further papers (after removing duplicates). After we read their full texts, we found 46 to satisfy our inclusion criteria. Among those 46, 15 were captured by the initial search. Tracking citations using Google Scholar provided 204 papers (after removing duplicates). After reading their full text, 49 were found to satisfy our inclusion criteria. Among those 49, 14 were captured by the initial search. The search results are outlined in Table 3. The final number of included papers was 93 (72 studies after grouping papers as described before). In total, the database search identified 36 papers while the other sources identified 57. Although the number of papers identified through other sources was large, many of them were variants of papers already included in the review.

The most common reasons for excluding papers on AQG were that the purpose of the generation was not related to education or there was no evaluation. Details of papers that were excluded after reading their full text are in the Appendix under "Excluded Studies".

Data Extraction Results

In this section, we provide our results and outline commonalities and differences with Alsubait's results (highlighted in the "Findings of Alsubait's Review" section).

Table 3 Sources used to obtain relevant papers and their contribution to the final results (* = after removing duplicates)

Source	Search results	No. included (based on title & abstract)	No. included (based on full text)
Computerised databases, journals, and conference proceedings			
ERIC	25	4	2
ACM	200	13	5
IEEE	107	34	13
INSPEC	174	58	24
Science direct	10	2	1
AIED (journal)	65	2	1
AIED (conference)	366	9	5
Total	974	122 (89 without duplicates)	51 (36 without duplicates)
Other sources			
Snowballing	–	169*	31
Google citation	–	204*	35
Other reviews	–	2	1
Ch and Saha (2018), Papasalouros and Chatzigiannakou (2018)			
Total (other sources)	–	375	67 (57 without duplicates)

The results are presented in the same order as our research questions. The main characteristics of the reviewed literature can be found in the Appendix under "Summary of Included Studies".

Rate of Publication

The distribution of publications by year is presented in Fig. 1. Putting this together with the results reported by Alsubait (2015), we notice a strong increase in publication starting from 2011. We also note that there were three workshops on QG[8] in 2008, 2009, and 2010, respectively, with one being accompanied by a shared task (Rus et al. 2012). We speculate that the increase starting from 2011 is because workshops on QG have drawn researchers' attention to the field, although the participation rate in the shared task was low (only five groups participated). The increase also coincides with the rise of MOOCs and the launch of major MOOC providers (Udacity, Udemy, Coursera and edX, which all started up in 2012 (Baturay 2015)) which provides another reason for the increasing interest in AQG. This interest was further boosted from 2015. In addition to the above speculations, it is important to mention that QG is closely related to other areas such as NLP and the Semantic Web. Being more mature and providing methods and tools that perform well have had an effect on the quantity and quality of research in QG. Note that these results are only related to question generation studies that focus on educational purposes and that there is a large volume of studies investigating question generation for other applications as mentioned in the "Search and Screening Results" section.

Types of Papers and Publication Venues

Of the papers published in the period covered by this review, conference papers constitute the majority (44 papers), followed by journal articles (32 papers) and workshop papers (17 papers). This is similar to the results of Alsubait (2015) with 34 conference papers, 22 journal papers, 13 workshop papers, and 12 other types of papers, including books or book chapters as well as technical reports and theses. In the Appendix, under "Publication Venues", we list journals, conferences, and workshops that published at least two of the papers included in either of the reviews.

Research Groups

Overall, 358 researchers are working in the area (168 identified in Alsubait's review and 205 identified in this review with 15 researchers in common). The majority of researchers have only one publication. In Appendix "Active Research Groups", we present the 13 active groups defined as having more than two publications in the period of both reviews. Of the 174 papers identified in both reviews, 64 were published by these groups. This shows that, besides the increased activities in the study of AQG, the community is also growing.

[8] http://www.questiongeneration.org/

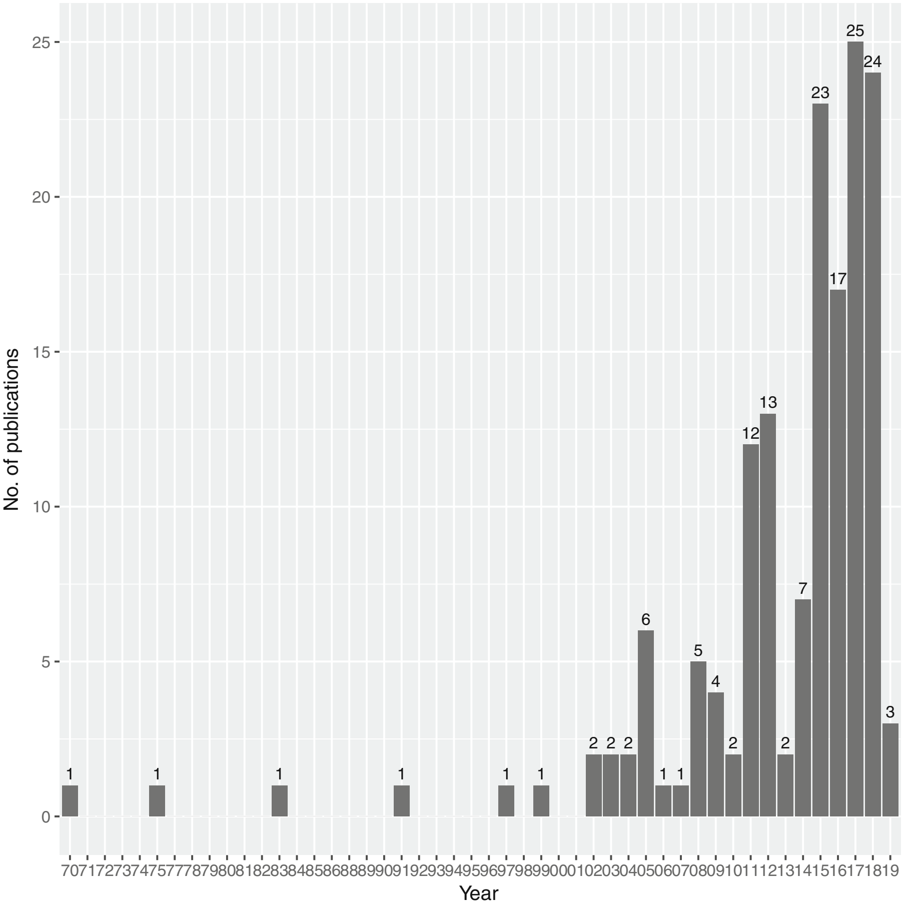

Fig. 1 Publications per year

Purpose of Question Generation

Similar to the results of Alsubait's review (Table 1), the main purpose of generating questions is to use them as assessment instruments (Table 4). Questions are also generated for other purposes, such as to be employed in tutoring or self-assisted learning systems. Generated questions are still used in experimental settings and only Zavala and Mendoza (2018) have reported their use in a class setting, in which the generator is used to generate quizzes for several courses and to generate assignments for students.

Generation Methods

Methods of generating questions have been classified in the literature (Yao et al. 2012) as follows: 1) syntax-based, 2) semantic-based, and 3) template-based.

Table 4 Purposes for automatically generating questions in the included studies. Note that a study can belong to more than one category

Purpose	No. of studies
Assessment	40
Education with no focus on a specific purpose	10
Self-directed learning, self-study or self-assessment	9
Learning support	9
Tutoring system or computer-assisted learning system	7
Providing practice questions	8
Providing questions for MOOCs or other courses	2
Active learning	1

Syntax-based approaches operate on the syntax of the input (e.g. syntactic tree of text) to generate questions. Semantic-based approaches operate on a deeper level (e.g. is-a or other semantic relations). Template-based approaches use templates consisting of fixed text and some placeholders that are populated from the input. Alsubait (2015) extended this classification to include two more categories: 4) rule-based and 5) schema-based. The main characteristic of rule-based approaches, as defined by Alsubait (2015), is the use of rule-based knowledge sources to generate questions that assess understanding of the important rules of the domain. As this definition implies that these methods require a deep understanding (beyond syntactic understanding), we believe that this category falls under the semantic-based category. However, we define the rule-based approach differently, as will be seen below. Regarding the fifth category, according to Alsubait (2015), schemas are similar to templates but are more abstract. They provide a grouping of templates that represent variants of the same problem. We regard this distinction between template and schema as unclear. Therefore, we restrict our classification to the template-based category regardless of how abstract the templates are.

In what follows, we extend and re-organise the classification proposed by Yao et al. (2012) and extended by Alsubait (2015). This is due to our belief that there are two relevant dimensions that are not captured by the existing classification of different generation approaches: 1) the level of understanding of the input required by the generation approach and 2) the procedure for transforming the input into questions. We describe our new classification, characterise each category and give examples of features that we have used to place a method within these categories. Note that these categories are not mutually exclusive.

- Level of understanding
  - Syntactic: Syntax-based approaches leverage syntactic features of the input, such as POS or parse-tree dependency relations, to guide question generation. These approaches do not require understanding of the semantics of the input in use (i.e. entities and their meaning). For example, approaches that select distractors based on their POS are classified as syntax-based.

  - Semantic: Semantic-based approaches require a deeper understanding of the input, beyond lexical and syntactic understanding. The information that these approaches use are not necessarily explicit in the input (i.e. they may require reasoning to be extracted). In most cases, this requires the use of additional knowledge sources (e.g., taxonomies, ontologies, or other such sources). As an example, approaches that use either contextual similarity or feature-based similarity to select distractors are classified as being semantic-based.

- Procedure of transformation

  - Template: Questions are generated with the use of templates. Templates define the surface structure of the questions using fixed text and placeholders that are substituted with values to generate questions. Templates also specify the features of the entities (either syntactic, semantic, or both), that can replace the placeholders.
  - Rule: Questions are generated with the use of rules. Rules often accompany approaches using text as input. Typically, approaches utilising rules annotate sentences with syntactic and/or semantic information. They then use these annotations to match the input to a pattern specified in the rules. These rules specify how to select a suitable question type (e.g. selecting suitable wh-words) and how to manipulate the input to construct questions (e.g. converting sentences into questions).
  - Statistical methods: This is where question transformation is learned from training data. For example, in Gao et al. (2018), question generation has been dealt with as a sequence-to-sequence prediction problem in which, given a segment of text (usually a sentence), the question generator forms a sequence of text representing a question (using the probabilities of co-occurrence that are learned from the training data). Training data has also been used in Kumar et al. (2015b) for predicting which word(s) in the input sentence is/are to be replaced by a gap (in gap-fill questions).

Regarding the level of understanding, 60 papers rely on semantic information and only ten approaches rely only on syntactic information. All except three of the ten syntactic approaches (Das and Majumder 2017; Kaur and Singh 2017; Kusuma and Alhamri 2018) tackle the generation of language questions. In addition, templates are more popular than rules and statistical methods, with 27 papers reporting the use of templates, compared to 16 and nine for rules and statistical methods, respectively. Each of these three approaches has its advantages and disadvantages. In terms of cost, all three approaches are considered expensive. Templates and rules require manual construction, while learning from data often requires a large amount of annotated data which is unavailable in many specific domains. Additionally, questions generated by rules and statistical methods are very similar to the input (e.g. sentences used for generation), while templates allow the generating of questions that differ from the surface structure of the input, in the use of words for example. However, questions generated from templates are limited in terms of their linguistic diversity. Note that some of the papers were classified as not having a method of transforming the input into questions because they only focused on distractor generation or gap-fill questions for which the stem is the same input statement with a word or a phrase being removed. Readers interested in studies that belong to a specific approach are referred to the "Summary of Included Studies" in the Appendix.

Generation Tasks

Tasks involved in question generation are explained below. We grouped the tasks into the stages of preprocessing, question construction, and post-processing. For each task, we provide a brief description, mention its role in the generation process, and summarise different approaches that have been applied in the literature. The "Summary of Included Studies" in the Appendix shows which tasks have been tackled in each study.

Preprocessing Two types of preprocessing are involved: 1) standard preprocessing and 2) QG-specific preprocessing. Standard preprocessing is common to various NLP tasks and is used to prepare the input for upcoming tasks; it involves segmentation, sentence splitting, tokenisation, POS tagging, and coreference resolution. In some cases, it also involves named entity recognition (NER) and relation extraction (RE). The aim of QG-specific preprocessing is to make or select inputs that are more suitable for generating questions. In the reviewed literature, three types of QG-specific preprocessing are employed:

- Sentence simplification: This is employed in some text-based approaches (Liu et al. 2017; Majumder and Saha 2015; Patra and Saha 2018b). Complex sentences, usually sentences with appositions or sentences joined with conjunctions, are converted into simple sentences to ease upcoming tasks. For example, Patra and Saha (2018b) reported that Wikipedia sentences are long and contain multiple objects; simplifying these sentences facilitates triplet extraction (where triples are used later for generating questions). This task was carried out by using sentence simplification rules (Liu et al. 2017) and relying on parse-tree dependencies (Majumder and Saha 2015; Patra and Saha 2018b).
- Sentence classification: In this task, sentences are classified into categories, which is, according to Mazidi and Tarau (2016a) and Mazidi and Tarau (2016b), a key to determining the type of question to be asked about the sentence. This classification was carried out by analysing POS and dependency labels, as in Mazidi and Tarau (2016a) and Mazidi and Tarau (2016b) or by using a machine learning (ML) model and a set of rules, as in Basuki and Kusuma (2018). For example, in Mazidi and Tarau (2016a, b), the pattern "S-V-acomp" is an adjectival complement that describes the subject and is therefore matched to the question template "Indicate properties or characteristics of S?"
- Content selection: As the number of questions in examinations is limited, the goal of this task is to determine important content, such as sentences, parts of sentences, or concepts, about which to generate questions. In the reviewed literature, the majority approach is to generate all possible questions and leave the

task of selecting important questions to exam designers. However, in some settings such as self-assessment and self-learning environments, in which questions are generated "on the fly", leaving the selection to exam designers is not feasible.

Content selection was of interest for those approaches that utilise text more than for those that utilise structured knowledge sources. Several characterisations of important sentences and approaches for their selection have been proposed in the reviewed literature which we summarise in the following paragraphs.

Huang and He (2016) defined three characteristics for selecting sentences that are important for reading assessment and propose metrics for their measurement: keyness (containing the key meaning of the text), completeness (spreading over different paragraphs to ensure that test-takers grasp the text fully), and independence (covering different aspects of text content). Olney et al. (2017) selected sentences that: 1) are well connected to the discourse (same as completeness) and 2) contain specific discourse relations. Other researchers have focused on selecting topically important sentences. To that end, Kumar et al. (2015b) selected sentences that contain concepts and topics from an educational textbook, while Kumar et al. (2015a) and Majumder and Saha (2015) used topic modelling to identify topics and then rank sentences based on topic distribution. Park et al. (2018) took another approach by projecting the input document and sentences within it into the same n-dimensional vector space and then selecting sentences that are similar to the document, assuming that such sentences best express the topic or the essence of the document. Other approaches selected sentences by checking the occurrence of, or measuring the similarity to, a reference set of patterns under the assumption that these sentences convey similar information to sentences used to extract patterns (Majumder and Saha 2015; Das and Majumder 2017). Others (Shah et al. 2017; Zhang and Takuma 2015) filtered sentences that are insufficient on their own to make valid questions, such as sentences starting with discourse connectives (e.g. thus, also, so, etc.) as in Majumder and Saha (2015).

Still other approaches to content selection are more specific and are informed by the type of question to be generated. For example, the purpose of the study reported in Susanti et al. (2015) is to generate "closest-in-meaning vocabulary questions"[9] which involve selecting a text snippet from the Internet that contains the target word, while making sure that the word has the same sense in both the input and retrieved sentences. To this end, the retrieved text was scored on the basis of metrics such as the number of query words that appear in the text.

With regard to content selection from structured knowledge bases, only one study focuses on this task. Rocha and Zucker (2018) used DBpedia to generate questions along with external ontologies; the ontologies describe educational standards according to which DBpedia content was selected for use in question generation.

Question Construction This is the main task and involves different processes based on the type of questions to be generated and their response format. Note that some

[9] Questions consisting of a text segment followed by a stem of the form: "The word X in paragraph Y is closest in meaning to:" and a set of options. See Susanti et al. (2015) for more details.

studies only focus on generating partial questions (only stem or distractors). The processes involved in question construction are as follows:

- Stem and correct answer generation: These two processes are often carried out together, using templates, rules, or statistical methods, as mentioned in the "Generation Methods" Section. Subprocesses involved are:
  - transforming assertive sentences into interrogative ones (when the input is text);
  - determination of question type (i.e. selecting suitable wh-word or template); and
  - selection of gap position (relevant to gap-fill questions).

- Incorrect options (i.e. distractor) generation: Distractor generation is a very important task in MCQ generation since distractors influence question quality. Several strategies have been used to generate distractors. Among these are selection of distractors based on word frequency (i.e. the number of times distractors appear in a corpus is similar to the key) (Jiang and Lee 2017), POS (Soonklang and Muangon 2017; Susanti et al. 2015; Satria and Tokunaga 2017a, b; Jiang and Lee 2017), or co-occurrence with the key (Jiang and Lee 2017). A dominant approach is the selection of distractors based on their *similarity* to the key, using different notions of similarity, such as syntax-based similarity (i.e. similar POS, similar letters) (Kumar et al. 2015b; Satria and Tokunaga 2017a, b; Jiang and Lee 2017), feature-based similarity (Wita et al. 2018; Majumder and Saha 2015; Patra and Saha 2018a, b; Alsubait et al. 2016; Leo et al. 2019), or contextual similarity (Afzal 2015; Kumar et al. 2015a, b; Yaneva and et al. 2018; Shah et al. 2017; Jiang and Lee 2017). Some studies (Lopetegui et al. 2015; Faizan and Lohmann 2018; Faizan et al. 2017; Kwankajornkiet et al. 2016; Susanti et al. 2015) selected distractors that are declared in a KB to be siblings of the key, which also implies some notion of similarity (siblings are assumed to be similar). Another approach that relies on structured knowledge sources is described in Seyler et al. (2017). The authors used query relaxation, whereby queries used to generate question keys are relaxed to provide distractors that share some of the key features. Faizan and Lohmann (2018) and Faizan et al. (2017) and Stasaski and Hearst (2017) adopted a similar approach for selecting distractors. Others, including Liang et al. (2017, 2018) and Liu et al. (2018), used ML-models to rank distractors based on a combination of the previous features.

  Again, some distractor selection approaches are tailored to specific types of questions. For example, for pronoun reference questions generated in Satria and Tokunaga (2017a, b), words selected as distractors do not belong to the same coreference chain as this would make them correct answers. Another example of a domain specific approach for distractor selection is related to gap-fill questions. Kumar et al. (2015b) ensured that distractors fit into the question sentence by calculating the probability of their occurring in the question.
- Feedback generation: Feedback provides an explanation of the correctness or incorrectness of responses to questions, usually in reaction to user selection. As feedback generation is one of the main interests of this review, we elaborate more fully on this in the "Feedback Generation" section.

- Controlling difficulty: This task focuses on determining how easy or difficult a question will be. We elaborate more on this in the section titled "Difficulty" .

Post-processing The goal of post-processing is to improve the output questions. This is usually achieved via two processes:

- Verbalisation: This task is concerned with producing the final surface structure of the question. There is more on this in the section titled "Verbalisation".
- Question ranking (also referred to as question selection or question filtering): Several generators employed an "over-generate and rank" approach whereby a large number of questions are generated, and then ranked or filtered in a subsequent phase. The ranking goal is to prioritise good quality questions. The ranking is achieved by the use of statistical models as in Blšták (2018), Kwankajornkiet et al. (2016), Liu et al. (2017), and Niraula and Rus (2015).

Input

In this section, we summarise our observations on which input formats are most popular in the literature published after 2014. One question we had in mind is whether structured sources (i.e. whereby knowledge is organised in a way that facilitates automatic retrieval and processing) are gaining more popularity. We were also interested in the association between the input being used and the domain or question types. Specifically, are some inputs more common in specific domains? And are some inputs more suitable for specific types of questions?

As in the findings of Alsubait (Table 1), text is still the most popular type of input with 42 studies using it. Ontologies and resource description framework (RDF) knowledge bases come second, with eight and six studies, respectively, using these. Note that these three input formats are shared between our review and Alsubit's review. Another input, used by more than one study, are question stems and keys, which feature in five studies that focus on generating distractors. See the Appendix "Summary of Included Studies" for types of inputs used in each study.

The majority of studies reporting the use of text as the main input are centred around generating questions for language learning (18 studies) or generating simple factual questions (16 studies). Other domains investigated are medicine, history, and sport (one study each). On the other hand, among studies utilising Semantic Web technologies, only one tackles the generation of language questions and nine tackle the generation of domain-unspecific questions. Questions for biology, medicine, biomedicine, and programming have also been generated using Semantic Web technologies. Additional domains investigated in Alsubait's review are mathematics, science, and databases (for studies using the Semantic Web). Combining both results, we see a greater variety of domains in semantic-based approaches.

Free-response questions are more prevalent among studies using text, with 21 studies focusing on this question type, 18 on multiple-choice, three on both free-response and multiple-choice questions, and one on verbal response questions. Some studies employ additional resources such as WordNet (Kwankajornkiet et al. 2016; Kumar et al. 2015a) or DBpedia (Faizan and Lohmann 2018; Faizan et al. 2017;

Tamura et al. 2015) to generate distractors. By contrast, MCQs are more prevalent in studies using Semantic Web technologies, with ten studies focusing on the generation of multiple-choice questions and four studies focusing on free-response questions. This result is similar to those obtained by Alsubait (Table 1) with free-response being more popular for generation from text and multiple-choice more popular from structured sources. We have discussed why this is the case in the "Findings of Alsubait's Review" Section.

Domain, Question Types and Language

As Alsubait found previously ("Findings of Alsubait's Review" section), language learning is the most frequently investigated domain. Questions generated for language learning target reading comprehension skills, as well as knowledge of vocabulary and grammar. Research is ongoing concerning the domains of science (biology and physics), history, medicine, mathematics, computer science, and geometry, but there are still a small number of papers published on these domains. In the current review, no study has investigated the generation of logic and analytical reasoning questions, which were present in the studies included in Alsubait's review. Sport is the only new domain investigated in the reviewed literature. Table 5 shows the number of papers in each domain and the types of questions generated for these domains (for more details, see the Appendix, "Summary of Included Studies"). As Table 5 illustrates, gap-fill and wh-questions are again the most popular. The reader is referred to the section "Findings of Alsubait's Review" for our discussion of reasons for the popularity of the language domain and the aforementioned question types.

With regard to the response format of questions, both free- and selected-response questions (i.e. MC and T/F questions) are of interest. In all, 35 studies focus on generating selected-response questions, 32 on generating free-response questions, and four studies on both. These numbers are similar to the results reported in Alsubait (2015), which were 33 and 32 papers on generation of free- and selected-response questions respectively (Table 1). However, which format is more suitable for assessment is debatable. Although some studies that advocate the use of free-response argue that these questions can test a higher cognitive level,[10] most automatically generated free-response questions are simple factual questions for which the answers are short facts explicitly mentioned in the input. Thus, we believe that it is useful to generate distractors, leaving to exam designers the choice of whether to use the free-response or the multiple-choice version of the question.

Concerning language, the majority of studies focus on generating questions in English (59 studies). Questions in Chinese (5 studies), Japanese (3 studies), Indonesian (2 studies), as well as Punjabi and Thai (1 study each) have also been generated. To ascertain which languages have been investigated before, we skimmed the papers identified in Alsubait (2015) and found three studies on generating questions in languages other than English: French in Fairon (1999), Tagalog in Montenegro et al.

[10] This relates to the processes required to answer questions as characterised in known taxonomies such as Bloom's taxonomy (Bloom et al. 1956), SOLO taxonomy (Biggs and Collis 2014) or Webb's depth of knowledge (Webb 1997).

(2012), and Chinese, in addition to English, in Wang et al. (2012). This reflects an increasing interest in generating questions in other languages, which possibly accompanies interest in NLP research in these domains. Note that there may be studies on other languages or more studies on the languages we have identified that we were not able to capture, because we excluded studies written in languages other than English.

Feedback Generation

Feedback generation concerns the provision of information regarding the response to a question. Feedback is important in reinforcing the benefits of questions especially in electronic environments in which interaction between instructors and students is limited. In addition to informing test takers of the correctness of their responses, feedback plays a role in correcting test takers' errors and misconceptions and in guiding them to the knowledge they must acquire, possibly with reference to additional materials.

This aspect of questions has been neglected in early and recent AQG literature. Among the literature that we reviewed, only one study, Leo et al. (2019), has generated feedback, alongside the generated questions. They generate feedback as a verbalisation of the axioms used to select options. In cases of distractors, axioms used to generate both key and distractors are included in the feedback.

We found another study (Das and Majumder 2017) that has incorporated a procedure for generating hints using syntactic features, such as the number of words in the key, the first two letters of a one-word key, or the second word of a two-words key.

Difficulty

Difficulty is a fundamental property of questions that is approximated using different statistical measures, one of which is *percentage correct* (i.e the percentage of examinees who answered a question correctly).[11] Lack of control over difficulty poses issues such as generating questions of inappropriate difficulty (inappropriately easy or difficult questions). Also, searching for a question with a specific difficulty among a huge number of generated questions is likely to be tedious for exam designers.

We structure this section around three aspects of difficulty models: 1) their generality, 2) features underlying them, and 3) evaluation of their performance.

Despite the growth in AQG, only 14 studies have dealt with difficulty. Eight of these studies focus on the difficulty of questions belonging to a particular domain, such as mathematical word problems (Wang and Su 2016; Khodeir et al. 2018), geometry questions (Singhal et al. 2016), vocabulary questions (Susanti et al. 2017a), reading comprehension questions (Gao et al. 2018), DFA problems (Shenoy et al. 2016), code-tracing questions (Thomas et al. 2019), and medical case-based questions (Leo et al. 2019; Kurdi et al. 2019). The remaining six focus on controlling the difficulty of non-domain-specific questions (Lin et al. 2015; Alsubait et al. 2016; Kurdi et al. 2017; Faizan and Lohmann 2018; Faizan et al. 2017; Seyler et al. 2017; Vinu and Kumar 2015a, 2017a; Vinu et al. 2016; Vinu and Kumar 2017b, 2015b).

[11] A percentage of 0 means that no one answered the question correctly (highly difficult question), while 100% means that everyone answered the question correctly (extremely easy question).

Table 5 Domains for which questions are generated and types of questions in the reviewed literature

Domain	No. of studies	Questions	No. of studies
Generic	34	Gap-fill questions	10
		Wh-questions	12
		What	7
		Where	6
		Who	5
		When, Why, How, and How many	4
		Which	2
		Whom, Whose, and How much	1
		Jeopardy-style questions	2
		Analogy	2
		Recognition, generalisation, and specification	1
		List and describe questions	1
		Summarise and name some	2
		Pattern-based questions	1
		Aggregation-based questions	1
		Definition	2
		Choose-the-type questions	1
		Comparison	1
		Description	1
		Not mentioned	1
		Other	3
Language learning	21	Gap-fill questions	8
		Wh-questions	4
		When	4
		What and Who	3
		Where and How many	2
		Which, Why, How, and How long	1
		TOEFL reference questions	1
		TOEFL vocabulary questions	1
		Word reading questions	1
		Vocabulary matching questions	1
		Reading comprehension (inference) questions	1
Biology	1	Input and output questions and function questions	1
		Inverse of the “feature specification” questions	1
		Wh-questions	1
		What and Where	1
History	1	Concept completion questions	1
		Casual consequence questions	1
		Composition questions	1
		Judgment questions	1
		Wh-questions (who)	1

Table 5 (continued)

Domain	No. of studies	Questions	No. of studies
Bio-medicine and Medicine	4	Case-based questions	2
		Definition	1
		Wh-questions	1
Geometry	1	Geometry questions	1
Physics	1		
Mathematics	4	Mathematical word problems	1
		Algebra questions	1
Computer science	3	Program tracing	1
		Deterministic finite automata (DFA) problems	1
		coding questions	1
Sport	1	Wh-questions	1

Table 6 shows the different features proposed for controlling question difficulty in the aforementioned studies. In seven studies, RDF knowledge bases or OWL ontologies were used to derive the proposed features. We observe that only a few studies account for the contribution of both stem and options to difficulty.

Difficulty control was validated by checking agreement between predicted difficulty and expert prediction in Vinu and Kumar (2015b), Alsubait et al. (2016), Seyler et al. (2017), Khodeir et al. (2018), and Leo et al. (2019), by checking agreement between predicted difficulty and student performance in Alsubait et al. (2016), Susanti et al. (2017a), Lin et al. (2015), Wang and Su (2016), Leo et al. (2019), and Thomas et al. (2019), by employing automatic solvers in Gao et al. (2018), or by asking experts to complete a survey after using the tool (Singhal et al. 2016). Expert reviews and mock exams are equally represented (seven studies each). We observe that the question samples used were small, with the majority of samples containing less than 100 questions (Table 7).

In addition to controlling difficulty, in one study (Kusuma and Alhamri 2018), the author claims to generate questions targeting a specific Bloom level. However, no evaluation of whether generated questions are indeed at a particular Bloom level was conducted.

Verbalisation

We define verbalisation as any process carried out to improve the surface structure of questions (grammaticality and fluency) or to provide variations of questions (i.e. paraphrasing). The former is important since linguistic issues may affect the quality of generated questions. For example, grammatical inconsistency between the stem and incorrect options enables test takers to select the correct option with no mastery of the required knowledge. On the other hand, grammatical inconsistency between the stem and the correct option can confuse test takers who have the required knowledge and would have been likely to select the key otherwise. Providing different phrasing for the question text is also of importance, playing a role in keeping test

Table 6 Features proposed for controlling the difficulty of generated questions

Reference	Feature
Lin et al. (2015)	Feature-based similarity between key and distractors
Singhal et al. (2015a, b, 2016)	Number and type of domain-objects involved
	Number and type of domain-rules involved
	User given scenarios
	Length of the solution
	Direct/indirect use of rules involved
Susanti et al. (2017a, b, 2015, 2016)	Reading passage difficulty
	Contextual similarity between key and distractors
	Distractor word difficulty level
Vinu and Kumar (2015a, 2017a),	Quality of hints (i.e. how much they reduce the answer space)
Vinu et al. (2016)	Popularity of predicates present in stems
and Vinu and Kumar (2017b)	Depth of concepts and roles present in a stem in class hierarchy
Vinu and Kumar (2015b)	Feature-based similarity between key and distractors
Alsubait et al. (2016)	Feature-based similarity between key and distractors
Kurdi et al. (2017)	
Shenoy et al. (2016)	Eight features specific to DFA problems such as the number of states
Wang and Su (2016)	Complexity of equations
	Presence of distraction (i.e. redundant information) in stem
Seyler et al. (2017)	Popularity of entities (of both question and answer)
	Popularity of semantic types
	Coherence of entity pairs (i.e. tendency to appear together)
	Answer type
Faizan and Lohmann (2018) and	Depth of the correct answer in class hierarchy
Faizan et al. (2017)	Popularity of RDF triples (of subject and object)
Gao et al. (2018)	Question word proximity hint (i.e. distance of all nonstop sentence words to the answer in the corresponding sentence)
Khodeir et al. (2018)	Number and types of included operators
	Number of objects in the story
Leo et al. (2019) and	Stem indicativeness
Kurdi et al. (2019)	Option entity difference
Thomas et al. (2019)	Number of executable blocks in a piece of code

takers engaged. It also plays a role in challenging test takers and ensuring that they have mastered the required knowledge, especially in the language learning domain. To illustrate, consider questions for reading comprehension assessment; if the questions match the text with a very slight variation, test takers are likely to be able to answer these questions by matching the surface structure without really grasping the meaning of the text.

Table 7 Types of evaluation employed for verifying difficulty models. An asterisk "*" indicates that no sufficient information about the reviewers is reported

	Type of evaluation		
Reference	Expert review	Mock exam	Other
Lin et al. (2015)		45 questions and 30 co-workers	
Singhal et al. (2015a, b, 2016)			10 experts
Susanti et al. (2015, 2016, 2017a, b)		120 questions and 88 participants	
Vinu and Kumar (2015a, 2017a, b) and Vinu et al. (2016)		24 questions and 54 students	
Vinu and Kumar (2015b)	31 questions and 7 reviewers		
Alsubait et al. (2016) and Kurdi et al. (2017)	115 questions and 3 reviewers	12 questions and 26 students	
Shenoy et al. (2016)		4 questions and 23 students	
Wang and Su (2016)		24 questions and 30 students	
Seyler et al. (2017)	150 questions and 13 reviewers*		
Faizan and Lohmann (2018) and Faizan et al. (2017)	14 questions and 50 reviewers*		
Gao et al. (2018)	200 questions and 5 reviewers*		2 automatic solvers
Khodeir et al. (2018)	25 questions and 4 reviewers		
Leo et al. (2019) and Kurdi et al. (2019)	435 questions and 15 reviewers	231 questions and 12 students	
Thomas et al. (2019)	36 questions and 12 reviewers*		

From the literature identified in this review, only ten studies apply additional processes for verbalisation. Given that the majority of the literature focuses on gap-fill question generation, this result is expected. Aspects of verbalisation that have been considered are pronoun substitutions (i.e. replacing pronouns by their antecedents) (Huang and He 2016), selection of a suitable auxiliary verb (Mazidi and Nielsen 2015), determiner selection (Zhang and VanLehn 2016), and representation of semantic entities (Vinu and Kumar 2015b; Seyler et al. 2017) (see below for more on this).

Other verbalisation processes that are mostly specific to some question types are the following: selection of singular personal pronouns (Faizan and Lohmann 2018; Faizan et al. 2017), which is relevant for Jeopardy questions; selection of adjectives for predicates (Vinu and Kumar 2017a), which is relevant for aggregation questions; and ordering sentences and reference resolution (Huang and He 2016), which is relevant for word problems.

For approaches utilising structured knowledge sources, semantic entities, which are usually represented following some convention such as using camel case (e.g anExampleOfCamelCase) or using underscore as a word separator, need to be represented in a natural form. Basic processing which includes word segmentation, adaptation of camel case, underscores, spaces, punctuation, and conversion of the segmented phrase into a suitable morphological form (e.g. "has pet" to "having pet"), has been reported in Vinu and Kumar (2015b). Seyler et al. (2017) used Wikipedia to verbalise entities, an entity-annotated corpus to verbalise predicates, and WordNet to verbalise semantic types. The surface form of Wikipedia links was used as verbalisation for entities. The annotated corpus was used to collect all sentences that contain mentions of entities in a triple, combined with some heuristic for filtering and scoring sentences. Phrases between the two entities were used as verbalisation of predicates. Finally, as types correspond to WordNet synsets, the authors used a lexicon that comes with WordNet for verbalising semantic types.

Only two studies (Huang and He 2016; Ai et al. 2015) have considered paraphrasing. Ai et al. (2015) employed a manually created library that includes different ways to express particular semantic relations for this purpose. For instance, "wife had a kid from husband" is expressed as "from husband, wife had a kid". The latter is randomly chosen from among the ways to express the marriage relation as defined in the library. The other study that tackles paraphrasing is Huang and He (2016) in which words were replaced with synonyms.

Evaluation

In this section, we report on standard datasets and evaluation practices that are currently used in the field (considering how QG approaches are evaluated and what aspects of questions such evaluation focuses on). We also report on issues hindering comparison of the performance of different approaches and identification of the best-performing methods. Note that our focus is on the results of evaluating the whole generation approach, as indicated by the quality of generated questions, and not on the results of evaluating a specific component of the approach (e.g. sentence selection or classification of question types). We also do not report on evaluations related to the usability of question generators (e.g. evaluating ease of use) or efficiency (i.e. time taken to generate questions). For approaches using ontologies as the main input, we consider whether they use existing ontologies or experimental ones (i.e. created for the purpose of QG), since Alsubait (2015) has concerns related to using experimental ontologies in evaluations (see "Findings of Alsubait's Review" section). We also reflect on further issues in the design and implementation of evaluation procedures and how they can be improved.

Standard Datasets In what follows, we outline publicly available question corpora, providing details about their content, as well as how they were developed and used in the context of QG. These corpora are grouped on the basis of the initial purpose for which they were developed. Following this, we discuss the advantages and limitations of using such datasets and call attention to some aspects to consider when developing similar datasets.

The identified corpora are developed for the following three purposes:

- Machine reading comprehension
    - The Stanford Question Answering Dataset (SQuAD)[12] (Rajpurkar et al. 2016) consists of 150K questions about Wikipedia articles developed by AMT co-workers. Of those, 100K questions are accompanied by paragraph-answer pairs from the same articles and 50K questions have no answer in the article. This dataset was used by Kumar et al. (2018) and Wang et al. (2018) to perform a comparison among variants of the generation approach they developed and between their approach and an approach from the literature. The comparison was based on the metrics BLEU-4, METEOR, and ROUGE-L which capture the similarity between generated questions and the SQuAD questions that serve as ground truth questions (there is more information on these metrics in the next section). That is, questions were generated using the 100K paragraph-answer pairs as input. Then, the generated questions were compared with the human-authored questions that are based on the same paragraph-answer pairs.
    - NewsQA[13] is another crowd-sourced dataset of about 120K question-answer pairs about CNN articles. The dataset consists of wh-questions and is used in the same way as SQuAD.
- Training question-answering (QA) systems
    - The 30M factoid question-answer corpus (Serban et al. 2016) is a corpus of questions automatically generated from Freebase.[14] Freebase triples (of the form: subject, relationship, object) were used to generate questions where the correct answer is the object of the triple. For example, the question: “What continent is bayuvi dupki in?” is generated from the triple (bayuvi dupki, contained by, europe). The triples and the questions generated from them are provided in the dataset. A sample of the questions was evaluated by 63 AMT co-workers, each of whom evaluated 44-75 examples; each question was evaluated by 3-5 co-workers. The questions were also evaluated by automatic evaluation metrics. Song and Zhao (2016a) performed a qualitative analysis comparing the

[12] This can be found at https://rajpurkar.github.io/SQuAD-explorer/

[13] This can be found at https://datasets.maluuba.com/NewsQA

[14] This is a collaboratively created knowledge base.

grammaticality and naturalness of questions generated by their approach and questions from this corpus (although the comparison is not clear).

- SciQ[15] (Welbl et al. 2017) is a corpus of 13.7K science MCQs on biology, chemistry, earth science, and physics. The questions target a broad cohort, ranging from elementary to college introductory level. The corpus was created by AMT co-workers at a cost of $10,415 and its development relied on a two-stage procedure. First, 175 co-workers were shown paragraphs and asked to generate questions for a payment of $0.30 per question. Second, another crowd-sourcing task in which co-workers validate the questions developed and provide them with distractors was conducted. A list of six distractors was provided by a ML-model. The co-workers were asked to select two distractors from the list and to provide at least one additional distractor for a payment of $0.20. For evaluation, a third crowd-sourcing task was created. The co-workers were provided with 100 question pairs, each pair consisting of an original science exam question and a crowd-sourced question in a random order. They were instructed to select the question likelier to be the real exam question. The science exam questions were identified in 55% of the cases. This corpus was used by Liang et al. (2018) to develop and test a model for ranking distractors. All keys and distractors in the dataset were fed to the model to rank. The authors assessed whether ranked distractors were among the original distractors provided with the questions.

- Question generation
  - The question generation shared task challenge (QGSTEC) dataset[16] (Rus et al. 2012) is created for the QG shared task. The shared task contains two challenges: question generation from individual sentences and question generation from a paragraph. The dataset contains 90 sentences and 65 paragraphs collected from Wikipedia, OpenLearn,[17] and Yahoo! Answers, with 180 and 390 questions generated from the sentences and paragraphs, respectively. A detailed description of the dataset, along with the results achieved by the participants, is given in Rus et al. (2012). Blšták and Rozinajová (2017, 2018) used this dataset to generate questions and compare their performance on correctness to the performance of the systems participating in the shared task.
  - Medical CBQ corpus (Leo et al. 2019) is a corpus of 435 case-based, auto-generated questions that follow four templates (“What is the most likely diagnosis?”, “What is the drug of choice?”, “What is the most likely clinical finding?”, and “What is the differential diagnosis?”). The

[15] Available at http://allenai.org/data.html

[16] The dataset can be obtained from https://github.com/bjwyse/QGSTEC2010/blob/master/QGSTEC-Sentences-2010.zip

[17] OpenLearn is an online repository that provides access to learning materials from The Open University.

questions are accompanied by experts' ratings of appropriateness, difficulty, and actual student performance. The data was used to evaluate an ontology-based approach for generating case-based questions and predicting their difficulty.

- MCQL is a corpus of about 7.1K MCQs crawled from the web, with an average of 2.91 distractors per question. The domains of the questions are biology, physics, and chemistry, and they target Cambridge O-level and college-level. The dataset was used in Blšták and Rozinajová (2017) to develop and evaluate a ML-model for ranking distractors.

Several datasets were used for assessing the ability of question generators to generate similar questions (see Table 8 for an overview). Note that the majority of these datasets were developed for purposes other than education and, as such, the educational value of the questions has not been validated. Therefore, while use of these datasets supports the claim of being able to generate human-like questions, it does not indicate that the generated questions are good or educationally useful. Additionally, restricting the evaluation of generation approaches to the criterion of being able to generate questions that are similar to those in the datasets does not capture their ability to generate other good quality questions that differ in surface structure and semantics.

Table 8 Information about question corpora that are used in the reviewed literature

Name	Size	Source	Development method	Content	Educationally relevant
The 30M Factoid Question Answer corpus	30M	Freebase	Automatic	Question-answer pairs (answer or triples)	no
SQuAD	150K	Wikipedia	Crowdsourcing	paragraph-answer pairs and questions generated based on the pairs	no
NewsQA	120K	CNN articles	Crowdsourcing	Question-answer pairs	no
SciQ	13.7K	Science study textbooks	Crowdsourcing	MCQs	yes
MCQL	7.1K	Web	Unknown	MCQs	no
QGSTEC dataset	570	Wikipedia, OpenLearn and Yahoo! Answers	Manual	Question-answer pairs (answer or paragraph)	no
Medical CBQ corpus	435	Elsevier's Merged Medical Taxonomy (EMMeT)	Automatic	MCQs	yes

Some of these datasets were used to develop and evaluate ML-models for ranking distractors. However, being written by humans does not necessarily mean that these distractors are good. This is, in fact, supported by many studies on the quality of distractors in real exam questions (Sarin et al. 1998; Tarrant et al. 2009; Ware and Vik 2009). If these datasets were to be used for similar purposes, distractors would need to be filtered based on their functionality (i.e. being picked by test takers as answers to questions).

We also observe that these datasets have been used in a small number of studies (1-2). This is partially due to the fact that many of them are relatively new. In addition, the design space for question generation is large (i.e. different inputs, question types, and domains). Therefore, each of these datasets is only relevant for a small set of question generators.

Types of Evaluation The most common evaluation approach is expert-based evaluation (n = 21), in which experts are presented with a sample of generated questions to review. Given that expert review is also a standard procedure for selecting questions for real exams, expert rating is believed to be a good proxy for quality. However, it is important to note that expert review only provides initial evidence for the quality of questions. The questions also need to be administered to a sample of students to obtain further evidence of their quality (empirical difficulty, discrimination, and reliability), as we will see later. However, invalid questions must be filtered first, and expert review is also utilised for this purpose, whereby questions indicated by experts to be invalid (e.g. ambiguous, guessable, or not requiring domain knowledge) are filtered out. Having an appropriate question set is important to keep participants involved in question evaluation motivated and interested in solving these questions.

One of our observations on expert-based evaluation is that only in a few studies were experts required to answer the questions as part of the review. We believe this is an important step to incorporate since answering a question encourages engagement and triggers deeper thinking about what is required to answer. In addition, expert performance on questions is another indicator of question quality and difficulty. Questions answered incorrectly by experts can be ambiguous or very difficult.

Another observation on expert-based evaluation is the ambiguity of instructions provided to experts. For example, in an evaluation of reading comprehension questions (Mostow et al. 2017), the authors reported different interpretations of the instructions for rating the overall question quality, whereby one expert pointed out that it is not clear whether reading the preceding text is required in order to rate the question as being of good quality. Researchers have also measured question acceptability, as well as other aspects of questions, using scales with a large number of categories (up to a 9-point scale) without a clear categorisation for each category. Zhang (2015) found that reviewers perceive scale differently and not all categories of scales are used by all reviewers. We believe that these two issues are reasons for low inter-rater agreement between experts. To improve the accuracy of the data obtained through expert review, researchers must precisely specify the criteria by which to evaluate questions. In addition, a pilot test needs to be conducted with experts to provide an opportunity for validating the instructions and ensuring that instructions and questions are easily understood and interpreted as intended by different respondents.

The second most commonly employed method for evaluation is comparing machine-generated questions (or parts of questions) to human-authored ones (n = 15), which is carried out automatically or as part of the expert review. This comparison is utilised to confirm different aspects of question quality. Zhang and VanLehn (2016) evaluated their approach by counting the number of questions in common between those that are human- and machine-generated. The authors used this method under the assumption that humans are likely to ask deep questions about topics (i.e. questions of higher cognitive level). On this ground, the authors claimed that an overlap means the machine was able to mimic this in-depth questioning. Other researchers have compared machine-generated questions with human-authored reference questions using metrics borrowed from the fields of text summarisation (ROUGE (Lin 2004)) and machine translation (BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005)). These metrics measure the similarity between two questions generated from the same text segment or sentence. Put simply, this is achieved by counting matching n-grams in the gold-standard question to n-grams in the generated question with some focusing on recall (i.e. how much of the reference question is captured in the generated question) and others focusing on precision (i.e. how much of the generated question is relevant). METEOR also considers stemming and synonymy matching. Wang et al. (2018) claimed that these metrics can be used as initial, inexpensive, large-scale indicators of the fluency and relevancy of questions. Other researchers investigated whether machine-generated questions are indistinguishable from human-authored questions by mixing both types and asking experts about the source of each question (Chinkina and Meurers 2017; Susanti et al. 2015; Khodeir et al. 2018). Some researchers evaluated their approaches by investigating the ability of the approach to assemble human-authored distractors. For example, Yaneva and et al. (2018) only focused on generating distractors given a question stem and key. However, given the published evidence of the poor quality of human-generated distractors, additional checks need to be performed, such as the functionality of these distractors.

Crowd-sourcing has also been used in ten of the studies. In eight of these, co-workers were employed to review questions while in the remaining three, they were employed to take mock tests. To assess the quality of their responses, Chinkina et al. (2017) included test questions to make sure that the co-workers understood the task and were able to distinguish low-quality from high-quality questions. However, including a process for validating the reliability of co-workers has been neglected in most studies (or perhaps not reported). Another validation step that can be added to the experimental protocol is conducting a pilot to test the capability of co-workers for review. This can also be achieved by adding validated questions to the list of questions to be reviewed by the co-workers (given the availability of a validated question set).

Similarly, students have been employed to review questions in nine studies and to take tests in a further ten. We attribute the low rate of question validation through testing with student cohorts to it being time-consuming and to the ethical issues involved in these experiments. Experimenters must ensure that these tests do not have an influence on students' grades or motivations. For example, if multiple auto-generated questions focus on one topic, students could perceive this as an important topic and pay more attention to it while studying for upcoming exams, possibly giving less attention to other topics not covered by the experimental exam. Difficulty of such experimental exams could also affect students. If an experimental test is very easy, students could expect upcoming exams to be the same, again paying less attention when studying for them. Another possible threat is a drop in student motivation triggered by an experimental exam being too difficult.

Finally, for ontology-based approaches, similar to the findings reported in the section "Findings of Alsubait's Review", most ontologies used in evaluations were hand-crafted for experimental purposes and the use of real ontologies was neglected, except in Vinu and Kumar (2015b), Leo et al. (2019), and Lopetegui et al. (2015).

Quality Criteria and Metrics Table 9 shows the criteria used for evaluating the quality of questions or their components. Some of these criteria concern the linguistic quality of questions, such as grammatical correctness, fluency, semantic ambiguity, freeness from errors, and distractor readability. Others are educationally oriented, such as educational usefulness, domain relevance, and learning outcome. There are also standard quality metrics for assessing questions, such as difficulty, discrimination, and cognitive level. Most of the criteria can be used to evaluate any type of question and only a few are applicable to a specific class of questions, such as the quality of blank (i.e. a word or a phrase that is removed from a segment of text) in gap-fill questions. As can be seen, human-based measures are the most common compared to automatic scoring and statistical procedures. More details about the measurement of these criteria and the results achieved by generation approaches can be found in the Appendix "Evaluation".

Performance of Generation Approaches and Gold Standard Performance

We started this systematic review hoping to identify standard performance and the best generation approaches. However, a comparison between the performances of various approaches was not possible due to heterogeneity in the measurement of quality and reporting of results. For example, scales that consist of different number of categories were used by different studies for measuring the same variables. We were not able to normalise these scales because most studies have only reported aggregated data without providing the number of observations in each rating scale category. Another example of heterogeneity is difficulty based on examinee performance. While some studies use percentage correct, others use Rasch difficulty without providing the raw data to allow the other metric to be calculated. Also, essential information that is needed to judge the trustability and generality of the results, such as sample size and selection method, was not reported in multiple studies. All of these issues preclude a statistical analysis of, and a conclusion about, the performance of generation approaches.

Quality Assessment Results

In this section, we describe and reflect on the state of experimental reporting in the reviewed literature.

Table 9 Evaluation metrics and number of papers that have used each metric

Metric	No. of studies
Question as a whole	
Statistical difficulty (i.e. based on examinee performance) and reviewer rating of difficulty	19
Question acceptability (often by domain experts)	17
Grammatical correctness	14
Semantic ambiguity	11
Educational usefulness (i.e. usability in a learning context)	10
Relevance to the input	8
Domain relevance	6
Fluency	6
Being indistinguishable from human-authored questions	6
ROUGE	6
BLEU	5
Overlap with human-authored questions	5
Discrimination	5
Freeness from errors	4
METEOR	3
Answerability	3
Cognitive level or depth	2
Learning outcome	2
Diversity of question types	2
How much the questions revealed about the answer	1
Options	
Distractor quality or plausibility	16
Answer correctness or distractor correctness	4
Distractor functionality (i.e. based on examinee performance)	2
Overlap with human-generated distractors	2
Distractor homogeneity	1
Option usefulness	1
Distractor matching intended type	1
Distractor readability	1
Stem	
Blank quality	3
Other	
Generality of the designed templates	1
Sentence quality	1

Overall, the experimental reporting is unsatisfactory. Essential information that is needed to assess the strength of a study is not reported, raising concerns about trustability and generalisability of the results. For example, the number of evaluated questions, the number of participants involved in evaluations, or both of these numbers are not mentioned in five, ten and five studies, respectively. Information about sampling strategy and how sample size was determined is almost never reported (see the Appendix, "Quality assessment").

A description of the participants' characteristics, whether experts, students, or co-workers, is frequently missing (neglected by 23 studies). Minimal information that needs to be reported about experts involved in reviewing questions, in addition to their numbers, is their teaching and exam construction experience. Reporting whether experts were paid or not is important for the reader to understand possible biases involved. However, this is not reported in 51 studies involving experiments with human subjects. Other additional helpful information to report is the time taken to review, because this would assist researchers to estimate the number of experts to recruit given a particular sample size, or to estimate the number of questions to sample given the available number of experts.

Characteristics of students involved in evaluations, such as their educational level and experience with the subject under assessment, are important for replication of studies. In addition, this information can provide a basis for combining evidence from multiple studies. For example, we could gain stronger evidence about the effect of specific features on question difficulty by combining studies investigating the same features with different cohorts. In addition, the characteristics of the participants are a possible justification for the difference in difficulty between studies. Similarly, criteria used for the selection of co-workers such as imposing a restriction on which countries they are from, or the number and accuracy of previous tasks in which they participated is important.

Some studies neglect to report on the total number of generated questions and the distribution of questions per categories (question types, difficulty levels, and question sources, when applicable), which are necessary to assess the suitability of sampling strategies. For example, without reporting the distribution of question types, making a claim based on random sampling that *"70% of questions are appropriate to be used in exams"* would be misleading if the distribution of question types is skewed. This is due to the sample not being representative of question types with a low number of questions. Similarly, if the majority of generated questions are easy, using a random sample will result in the underrepresentation of difficult questions, consequently precluding any conclusion about difficult questions or any comparison between easy and difficult questions.

With regard to measurement descriptions, 10 studies fail to report information sufficient for replication, such as instructions given to participants and a description of the rating scales. Another limitation concerning measurements is the lack of assessment of inter-rater reliability (not reported by 43 studies). In addition, we observed a lack of justification for experimental decisions. Examples of this are the sources from which questions were generated, when particular texts or knowledge sources were selected without any discussion of whether these sources were representative and of what they were representative. We believe that generation challenges and question quality issues that might be encountered when using different sources need to be raised and discussed.

Conclusion and Future Work

In this paper, we have conducted a comprehensive review of 93 papers addressing the automatic generation of questions for educational purposes. In what follows, we summarise our findings in relation to the review objectives.

Providing an Overview of the AQG Community and its Activities

We found that AQG is an increasing activity of a growing community. Through this review, we identified the top publication venues and the active research groups in the field, providing a connection point for researchers interested in the field.

Summarising Current QG Approaches

We found that the majority of QG systems focus on generating questions for the purpose of assessment. The template-based approach was the most common method employed in the reviewed literature. In addition to the generation of complete questions or of question components, a variety of pre- and post-processing tasks that are believed to improve question quality have been investigated. The focus was on the generation of questions from text and for the language domain. The generation of both multiple-choice and free-response questions was almost equally investigated with a large number of studies focusing on wh-word and gap-fill questions. We also found increased interest in generating questions in languages other than English. Although extensive research has been carried out on QG, only a small proportion of these tackle the generation of feedback, verbalisation of questions, and the control of question difficulty.

Identifying Gold Standard performance in AQG

Incomparability of the performance of generation approaches is an issue we identified in the reviewed literature. This issue is due to the heterogeneity in both measurement of quality and reporting of results. We suggest below how the evaluation of questions and reporting of results can be improved to overcome this issue.

Tracking the Evolution of AQG Since Alsubait's Review

Our results are consistent with the findings of Alsubait (2015). Based on these findings, we suggest that research in the area can be extended in the following directions (starting at the question level before moving on to the evaluation and research in closely related areas):

Improvement at the Question Level

Generating Questions with Controlled Difficulty As mentioned earlier, there is little research on question difficulty and what there is mostly focuses on either stem or distractor difficulty. The difficulty of both stem and options plays a role in overall difficulty and therefore needs to be considered together and not in isolation. Furthermore, controlling MCQ difficulty by varying the similarity between key and distractors is a common feature found in multiple studies. However, similarity is only one facet of difficulty and there are others that need to be identified and integrated into the generation process. Thus, the formulation of a theory behind an intelligent automatic question generator capable of both generating questions and accurately controlling their difficulty is at the heart of AQG research. This would be used for improving the quality of generated questions by filtering inappropriately easy or difficult questions which is especially important given the large number of questions.

Enriching Question Forms and Structures One of the main limitations of existing works is the simplicity of generated questions, which has also been highlighted in Song and Zhao (2016b). Most generated questions consist of a few terms and target lower cognitive levels. While these questions are still useful, there is a potential for improvement by exploring the generation of other, higher order and more complex, types of questions.

Automating Template Construction The template library is a major component of question generation systems. At present, the process of template construction is largely manual. The templates are either developed through analysing a set of handwritten questions manually or through consultation with domain experts. While one of the main motivations for generating questions automatically is cost reduction, both of these template acquisition techniques are costly. In addition, there is no evidence that the set of templates defined by a few experts is typical of the set of questions used in assessments. We attribute part of the simplicity of the current questions to the cost, both in terms of time and resources, of both template acquisition techniques.

The cost of generating questions automatically could be reduced further by automatically constructing templates. In addition, this would contribute to the development of more diverse questions.

Verbalisation Employing natural language generation and processing techniques in order to present questions in natural and correct forms and to eliminate errors that invalidate questions, such as syntactic clues, are important steps to take before questions can be used beyond experimental settings for assessment purposes.

Feedback Generation As has been seen in both reviews, work on feedback generation is almost non-existent. Developing mechanisms for producing rich, effective feedback is one of the features that needs to be integrated into the generation process. This includes different types of feedback, such as formative, summative, interactive, and personalised feedback.

Improvement of Evaluation Methods

Using Human-Authored Questions for Evaluation Evaluating question quality, whether by means of expert review or mock exams, is an expensive and time consuming process. Analysing existing exam performance data is a potential source for evaluating question quality and difficulty prediction models. Translating human-authored questions to a machine-processable representation is a possible method for evaluating the ability of generation approaches to generate human-like questions. Regarding the evaluation of difficulty models, this can be done by translating questions to a machine-processable representation, computing the features of these questions, and examining their effect on difficulty. This analysis also provides an understanding of pedagogical content knowledge (i.e. concepts that students often find difficult and usually have misconceptions about). This knowledge can be integrated into difficulty prediction models, or used for question selection and feedback generation.

Standardisation and Development of Automatic Scoring Procedures To ease comparison between different generation approaches, which was difficult due to heterogeneity in measurement and reporting as well as ungrounded heterogeneity needs to be eliminated. The development of standard and well defined scoring procedures is important to reduce heterogeneity and improve inter-rater reliability. In addition, developing automatic scoring procedures that correlate with human ratings are also important since this will reduce evaluation cost and heterogeneity.

Improvement of Reporting We also emphasise the need for good experimental reporting. In general, authors should improve reporting on their generation approaches and on evaluation, which are both essential for other researchers who wish to compare their approaches with existing approaches. At a minimum, data extracted in this review (refer to questions under OBJ2 and OBJ3) should be reported in all publications on AQG. To ensure quality, journals can require authors to be complete a checklist prior to peer review, which has shown to improve the reporting quality (Han et al. 2017). Alternatively, text-mining techniques can be used for assessing the reporting quality by targeting key information in AQG literature, as has been proposed in Flórez-Vargas et al. (2016).

Other Areas of Improvement and Further Research

Assembling Exams from the Generated Questions Although there is a large amount of work that needs to be done at the question level before moving to the exam level, further work in extending the difficulty models, enriching question form and structure, and improving presentation are steps towards this goal. Research in these directions will open new opportunities for AQG research to move towards assembling exams automatically from generated questions. One of the challenges in exam generation is the selection of a question set that is of appropriate difficulty with good coverage of the material. Ensuring that questions do not overlap or provide clues for other questions also needs to be taken into account. The AQG field could adopt ideas from the question answering field in which question entailment has been investigated (for example, see the work of Abacha and Demner-Fushman (2016)). Finally, ordering questions in a way that increases motivation and maximises the accuracy of scores is another interesting area.

Mining Human-Authored Questions While existing researchers claim that the questions they generate can be used for educational purposes, these claims are not generally supported. More attention needs to be given to the educational value of generated questions.

In addition to potential use in evaluation, analysing real, good quality exams can help to gain insights into what questions need to be generated so that the generation addresses real life educational needs. This will also help to quantify the characteristics of real questions (e.g. number of terms in real questions) and direct attention to what needs to be done and where the focus should be in order to move to exam generation. Additionally, exam questions reflect what should be included in similar assessments that, in turn, can be further used for content selection and the ranking of questions. For example, concepts extracted from these questions can inform the selection of existing textual or structured sources and the quantifying of whether or not the contents are of educational relevance.

Other potential advantages that the automatic mining of questions offers are the extraction of question templates, a major component of automatic question generators, and improving natural language generation. Besides, mapping the information contained in existing questions to an ontology permits modification of these questions, prediction of their difficulty, and the formation of theories about different aspects of the questions such as their quality.

Similarity Computation and Optimisation A variety of similarity measures have been used in the context of QG to select content for questions, to select plausible distractors and to control question difficulty (see "Generation Tasks" section for examples). Similarity can also be employed in suggesting a diverse set of generated questions (i.e. questions that do not entail the same meaning regardless of their surface structure). Improving computation of the similarity measures (i.e. speed and accuracy) and investigating other types of similarity that might be needed for other question forms are all considered sidelines that have direct implications for improving the current automatic question generation process. Evaluating the performance of existing similarity measures in comparison to each other and whether or not cheap similarity measures can approximate expensive ones are further interesting objects of study.

Source Acquisition and Enrichment As we have seen in this review, structured knowledge sources have been a popular source for question generation, either by themselves or to complement texts. However, knowledge sources are not available for many domains, while those that are developed for purposes other than QG might not be rich enough to generate good quality questions. Therefore, they need to be adapted or extended before they can be used for QG. As such, investigating different approaches for building or enriching structured knowledge sources and gaining further evidence for the feasibility of obtaining good quality knowledge sources that can be used for question generation, are crucial ingredients for their successful use in question generation.

Limitations

A limitation of this review is the underrepresentation of studies published in languages other than English. In addition, ten papers were excluded because of the unavailability of their full texts.

Appendix

Search Queries

Table 10 Details of search terms used

Database	Search query	Filter
ERIC	abstract: “question generation” pubyear:2015	–
	abstract: “question generation” pubyear:2016	–
	abstract: “question generation” pubyear:2017	–
	abstract: “question generation” pubyear:2018	–
	abstract: “question generation” pubyear:2019	–
ACM	question generation	Publication year $\geq$ 2015 Abstract search
IEEE	question NEAR/5 generation	Year: 2015 - 2019
INSPEC	question NEAR generation	Year: 2015 - 2019
Science direct	“question generation”	Year: 2015 - 2019 Title, abstract, keywords
AIED	–	Year: 2015, 2017 and 2018

Excluded Studies

Table 11 Number of excluded papers published between 2015 and 2018 and reasons for their exclusion

Reason for exclusion	No.
Purpose is not education	91
No evaluation of generated questions	39
Purpose is not clear	19
Not peer reviewed	14
Extension on a paper before 2014 and no significant change made to the system	10
No full text available	10
Selection from a question bank and no question generation	9
Not in English	9
No sufficient description of how questions are generated	7
The QG approach is based on substitution of placeholders with values from a predefined set	5
Review paper	2

Publication Venues

Table 12 Top publishing venues of AQG papers

Name	No. of papers
Journals	
1. Dialogue and Discourse	3
2. IEEE Transactions on Learning Technologies	3
3. Natural Language Engineering	3
4. Research and Practice in Technology Enhanced Learning	3
Conferences	
5. Artificial Intelligence in Education	7
6. IEEE International Conference on Advanced Learning Technologies	3
7. International Conference on Intelligent Tutoring Systems	3
8. IEEE International Conference on Cognitive Infocommunications	2
9. IEEE International Conference on Semantic Computing	2
10. IEEE International Conference on Tools with Artificial Intelligence	2
11. IEEE TENCON	2
12. The International Conference on Computer Supported Education	2
13. The International Joint Conference on Artificial Intelligence	2
Workshops and other venues	
14. The Workshop on Innovative Use of NLP for Building Educational Applications	12
15. The Workshop on Building Educational Applications Using NLP	4
16. OWL: Experiences and Directions (OWLED)	2
17. The ACM Technical Symposium on Computer Science Education	2
18. The Workshop on Natural Language Processing Techniques for Educational Applications	2
19. The Workshop on Question Generation	2

Active Research Groups

Table 13 Research groups with more than two publications in AQG (ordered by number of publications)

Authors	Affiliation, Country	Publications
1. T. Alsubait, G. Kurdi, J. Leo, N. Matentzoglu, B. Parsia and U. Sattler	The University of Manchester, UK	(Alsubait et al. 2012a, b, c, 2013, 2014a, b, 2016; Kurdi et al. 2017, 2019; Leo et al. 2019)
2. Y. Hayashi, C. Jouault and K. Seta	Osaka Prefecture University, Japan	(Jouault and Seta 2014; Jouault et al. 2015a, b, 2016a, b, 2017)
3. Y. Huang	National Taiwan University, Taiwan	(Mostow et al. 2004; Beck et al. 2004; Mostow and Chen 2009; Huang et al. 2014; Huang and Mostow 2015; Mostow et al. 2017)
J. Beck, J. Bey, W. Chen, A. Cuneo, D. Gates, H. Jang, J. Mostow, J. Sison, B. Tobin, J. Valeri and A. Weinstein	Carnegie Mellon University, USA	
M. C. Chen, Y. S. Sun and Y. Tseng	Unknown	
4. L. Liu	Chongqing University, China	(Liu et al. 2012a, b, 2014, 2017, 2018; Liu and Calvo 2012)
M. Liu	Southwest University, China	
V. Rus	University of Memphis, USA	
A. Aditomo, R. Calvo and L. Augusto Pizzato	University of Sydney, Australia	
5. N. Afzal, L. Ha and R. Mitkov	University of Wolverhampton, UK	(Mitkov and Ha 2003; Mitkov et al. 2006; Afzal et al. 2011; Afzal and Mitkov 2014; Afzal 2015)
A. Farzindar	NLP Technologies Inc, Canada	
6. V. Ellampallil Venugopal and P. Kumar	Indian Institute of Technology Madras, India	(Vinu and Kumar 2015a, b, 2017a, b; Vinu et al. 2016)
7. K. Mazidi, R. Nielsen and P. Tarau	University of North Texas, USA	(Mazidi and Nielsen 2014, 2015; Mazidi and Tarau 2016a, b; Mazidi 2018)
8. R. Goyal, M. Henz and R. Singhal	National University of Singapore, Singapore	(Singhal and Henz 2014; Singhal et al. 2015a, b; Singhal et al. 2016)
9. M. Heilman and N. A. Smith	Carnegie Mellon University, Pennsylvania	(Heilman and Smith 2009, 2010a, b; Heilman 2011)
10. H. Nishikawa, Y. Susanti and T. Tokunaga,	Tokyo Institute of Technology, Japan	(Susanti et al. 2015; Susanti et al. 2016; Susanti et al. 2017a; 2017b)
R. Iida	National Institute of Information and Communication Technology, Japan	
H. Obari	Aoyama Gakuin University, Japan	

Table 13 (continued)

Authors	Affiliation, Country	Publications
11. L. Bednarik, L. Kovacs and G. Szeman	University of Miskolc, Hungary	(Bednarik and Kovacs 2012a b; Kovacs and Szeman 2013)
12. M. Blšták and V. Rozinajová	Slovak University of Technology in Bratislava, Slovakia	(Blšták and Rozinajová 2017; Blšták 2018; Blšták and Rozinajová 2018)
13. M. Majumder	Vidyasagar University, India	(Majumder and Saha, 2015;
S. Patra and S. Saha	Birla Institute of Technology Mesra, India	Patra and Saha 2018b; 2018a)

Summary of Included Studies

Table 14 Basic information about the reviewed literature (ordered by publication year). Verb. = verbalisation; language = language of questions; gen. = generation; NR = not reported, and NC = not clear

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
1. Afzal (2015)	assessment	text	text corpus; untagged word patterns; PoS-tagged word patterns; verb-centred patterns; transformation rules	generic	wh	MC	English	no	no	no	expert review
2. Ai et al. (2015)	assessment	text	lexico-syntactic patterns	language (RC)	Which one of the following four facts can be inferred from the text?	MC	English	no	no	yes	student review
3. Fattoh et al. (2015)	education	text	–	generic	wh	FR	English	no	no	no	automatic evaluation
4. Huang and Mostow (2015) and Mostow et al. (2017)	assessment	text	–	language (RC)	gap-fill	MC	English	no	no	no	expert review; mock exam (with students); comparison with human-authored questions

Table 14 1 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
5. Kumar et al. (2015b)	self-assessment	text	text corpus	generic	gap-fill	MC	English	no	no	no	student review
6. Kumar et al. (2015a)	self-learning and self-assessment; practice questions	text	WordNet; text corpus	generic	gap-fill	MC	English	no	no	no	crowdsourcing review
7. Lin et al. (2015)	education	RDF KB	–	generic	NR	MC	English	yes	no	no	mock exam (crowdsourcing)
8. Lopetegui et al. (2015)	support learning	ontology	–	biomedicine	definition	MC	English	no	no	no	expert review
9. Majumder and Saha (2015)	assessment; active learning	text	gazetteer lists	sport	wh	MC	English	no	no	no	review (not clear by who)
10. Mazidi and Nielsen (2015)	tutoring	text	–	generic	wh	FR	English	no	no	yes	crowdsourcing review
11. Niraula and Rus (2015)	assessment	question	annotations of question quality	generic	gap-fill	FR	English	no	no	no	automatic evaluation
12. Odilinye et al. (2015)	self-learning	text	list of concepts; WordNet	generic	NR	FR	English	no	no	no	review (not clear by who); comparison with human-authored questions

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
13. Polozov et al. (2015)	education	requirement	ontology; verbalisation templates	math	word problem	FR	English	no	no	yes	crowdsourcing review; mock exam (crowdsourcing); comparison with human-authored questions
14. Shirude et al. (2015)	assessment; providing questions for academic courses and school textbooks	data table	–	generic	multiple types of questions	FR	English	no	no	no	comparison with human-authored distractors
15. Singhal et al. (2015a, b, 2016)	assessment; providing practice questions	first order logic formulas	–	physic and geometry	figure scenario questions	FR	NA	yes	no	no	expert review
16. Susanti et al. (2015, 2016, 2017a, b)	providing practice questions; self-study; assessment	target word with its POS and a word sense; WordNet	word frequency list; JACET8000	language	closest-in-meaning vocabulary questions	MC	English	yes	no	no	expert review; comparison with human-authored questions; mock exam (with students)
17. Tamura et al. (2015)	tutoring	text	RDF KB (DBpedia)	history	wh	MC	Japanese	no	no	no	review (not clear by who)

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
18. Vinu and Kumar (2015a, 2017a, b and Vinu et al. (2016))	assessment	ontology	–	generic	pattern-based questions; aggregation-based questions	MC	English	yes	no	yes	mock exam (with students); comparison with human-authored questions
19. Vinu and Kumar (2015b)	assessment	ontology	templates	generic	factual questions	MC	English	yes	no	yes	expert review
20. Zhang and Takuma (2015)	computer-assisted learning system	text	Kanji list database and word list database	language (reading)	read word in sentence	sound	Japanese	no	no	no	student review
21. Alsubait et al. (2016) and Kurdi et al. (2017)	assessment	ontology	–	generic	definition; recognition; specification; specification2; generalisation1; generalisation2; analogy	MC	English	yes	no	no	expert review; mock exam (with students); author review
22. Araki et al. (2016)	assessment	text	manual annotations of the input text	language (RC)	12 type of questions	MC	English	no	no	no	expert review
23. Hill and Simha (2016)	support learning	text	WordNet; Google Books n-grams Corpus	language (RC)	gap-fill	MC	English	no	no	no	review (not clear by who)

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
24. Huang and He (2016)	assessment	text	–	language (RC)	short answer questions	FR	English	no	no	yes	expert review; mock exam (with students); comparison with another generator; comparison with human-authored questions
25. Jouault et al. (2015a, b, 2016a, b, 2017)	support learning	RDF KB	–	history	wh	FR	English	no	no	no	expert review; comparison with human-authored questions; student review
26. Kwankajornkiet et al. (2016)	assessment	text	WordNet dictionary	language (RC)	gap-fill	MC	Thai	no	no	no	review (not clear by who)
27. Mazidi and Tarau (2016a, b)	tutoring	text	syntactic patterns	generic	wh	FR	English	no	no	no	crowdsourcing review; comparison with another generator
28. Shenoy et al. (2016)	assessment; practice	question		computer science	DFA problem	FR	English	yes	no	NC	mock exam (with students)
29. Song and Zhao (2016a, b)	assessment	RDF KB	annotation	generic	wh; other	FR	English	no	no	no	automatic evaluation

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
30. Wang and Su (2016)	providing practice questions	NC		math	word problem	FR	English	yes	no	no	mock exam (with students); student review
31. Zhang and VanLehn (2016)	assessment	ontology	–	biology	what is questions; input and output questions; where questions; function questions	FR	English	no	no	yes	student review; comparison with human-authored questions
32. Adithya and Singh (2017)	self-assessment	text	POS patterns	generic	questions about number, location, and name of a person	MC	English	no	no	no	NC
33. Blšták and Rozinajová (2017) and Blšták (2018)	self-learning	text	–	generic	wh and true-false	FR	English	no	no	no	comparison with another generator; review (not clear by who); comparison with human-authored questions
34. Chinkina et al. (2017)	support learning	text	–	language	wh; gap-fill	FR	English	no	no	no	crowdsourcing review
35. Chinkina and Meurers (2017)	support learning	text	–	language (linguistic forms and grammars)	form exposure questions; grammar-concept questions	FR	English	no	no	no	crowdsourcing review

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
36. Das and Majumder (2017)	assessment	text	–	generic	gap-fill	FR	English	no	no	no	mock exam (with students)
37. Gupta et al. (2017)	education	question	text	math	word problem	FR	English	no	no	no	student review
38. Jiang and Lee (2017)	education	question stem; question key	wiki corpus	language (vocabulary learning)	gap-fill	MC	Chinese	no	no	no	expert review
39. Kaur and Singh (2017)	education	text	–	NC	wh	FR	Punjabi	no	no	no	NC
40. Liang et al. (2017)	assessment	question stem; question key	question corpus	generic	gap-fill	MC	English	no	no	no	automatic evaluation; comparison with another generator
41. Liu et al. (2017)	assessment; tutoring; support learning	text	patterns	language (RC)	wh	FR	Chinese	no	no	no	automatic evaluation
42. Olney et al. (2017)	education	text	list of 1,000 most frequent words of English	language (RC)	gap-fill	MC	English	no	no	no	mock exam (crowdsourcing)
43. Santhanavijayan et al. (2017)	assessment	text	search query	generic	gap-fill; analogy	MC	NC	no	no	no	NC

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
44. Satria and Tokunaga (2017a, b)	providing practice questions	text	rules	language	pronoun reference questions	MC	English	no	no	no	author review; mock exam (with students); expert review; comparison with human-authored questions
45. Seyler et al. (2017)	self-learning and self-assessment; assessment	RDF KB	annotated text corpus; WordNet; question corpus annotated with difficulty	generic	Jeopardy questions	MC	English	yes	no	yes	automatic evaluation; crowdsourcing review
46. Shah et al. (2017)	tutoring; self-assessment; MOOC	text	list of Wikipedia links	generic	gap-fill	MC	English	no	no	no	expert review
47. Soonklang and Muangon (2017)	assessment	text	–	language	gap-fill; error correction	MC; T/F; FR	English	no	no	no	NC
48. Stasaski and Hearst (2017)	assessment	ontology	–	generic	specification	MC	English	no	no	no	expert review
49. Basuki and Kusuma (2018)	assessment	text	patterns	generic	wh	FR	Indonesian	no	no	no	NC

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
50. Blšták and Rozinajová (2018)	assessment	text	questions	generic	wh	FR	English	no	no	no	comparison with another generator; comparison with human-authored questions
51. Faizan and Lohmann (2018) and Faizan et al. (2017)	assessment	text	RDF KB (DBpedia and YAGO)	generic	gap-fill; choose-the-type questions; Jeopardy questions	MC	English	yes	no	yes	crowdsourcing review
52. Flor and Riordan (2018)	education	text	–	generic	wh; yes/no questions	FR;T/F	English	no	no	no	expert review; comparison with another generator
53. Gao et al. (2018)	education	text; question key	–	language (RC)	wh	FR	English	yes	no	no	review (not clear by who); automatic evaluation; comparison with another generator
54. Killawala et al. (2018)	assessment	text	question corpus; NLTK dictionary and WordNet	generic	wh; true/false; gap-fill	FR and MC	English	no	no	no	student review; comparison with human-authored questions

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
55. Khodeir et al. (2018)	tutoring	NC	–	math	word problem	FR	English	yes	no	yes	expert review; student review; comparison with human-authored questions
56. Kumar et al. (2018)	assessment	text	–	language (RC)	wh	FR	English	no	no	no	automatic evaluation; expert review
57. Kusuma and Alhamri (2018)	assessment	text	patterns	generic	NC	FR	Indonesian	no	no	no	expert review
58. Lee et al. (2018)	support learning	text	annotation of question quality	language (RC)	wh	FR	Chinese	no	no	no	NC
59. Liang et al. (2018)	assessment	question stem; question key; distractor set to rank (in ranking MC)	question corpus	generic	NR	MC	English	no	no	no	automatic evaluation
60. Liu et al. (2018)	assessment	text	annotations; Hownet	language	vocabulary related (seems gap-fill)	MC	Chinese	no	no	no	mock exam (with students)
61. Marrese-Taylor et al. (2018)	self-learning	text	question corpus	language	gap-fill	FR	English	no	no	no	automatic evaluation

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
62. Mazidi (2018)	education	text	–	generic	summary, comparison, description and definition questions	FR	English	no	no	no	crowdsourcing review; comparison with another generator; comparison with human-authored questions
63. Park et al. (2018)	assessment; practice questions	text	–	generic	gap-fill	MC	English	no	no	no	student review
64. Patra and Saha (2018a, b)	assessment	question stem; question key	Wikipedia; WordNet	sport	NR	MC	English	no	no	no	expert review; comparison with human-authored distractors; comparison with another generator
65. Rocha and Zucker (2018)	assessment; support learning	RDF KB	OWL ontology	generic	NR	MC	English	no	no	no	expert review
66. Wang et al. (2018)	assessment	text; question key	annotation	generic	wh	FR	English	no	no	no	automatic evaluation; expert review
67. Wita et al. (2018)	assist vocabulary learning process	vocabulary semantic graph	Japanese WordNet	language (vocabulary learning)	vocabulary matching questions	MC	Japanese	no	no	no	expert review

Table 14 (continued)

Reference	Purpose	Input	Additional input	Domain	Question format	Response format	Language	Difficulty	Feedback gen.	Verb.	Evaluation
68. Yaneva and et al. (2018)	assessment	question stem; question key	embedding vectors	medicine	case-based questions	MC	English	no	no	no	automatic evaluation
69. Zhang et al. (2018)	assessment	text, ontology and term base	–	medicine	wh; yes/no questions	FR and MC	Chinese	no	no	no	student review
70. Zavala and Mendoza (2018)	assessment	RDF KB	–	computer science (programming)	coding questions	FR	English	no	no	no	student study (post-pre test design)
71. Leo et al. (2019) and Kurdi et al. (2019)	assessment	ontology	–	medicine	case-based questions	MC	English	yes	yes	no	expert review; mock exam (with students)
72. Thomas et al. (2019)	providing practice questions	pattern	–	computer science	program tracing	MC	English	yes	no	no	review (not clear by who)

Table 15 Classification of approaches used in the included studies (ordered by publication year). Note that a study can appear in more than one category. NA = not applicable and NC = not clear

	Understanding level		**Question formation method**			
	Syntax based	**Semantic based**	**Template based**	**Rule based**	**Statistical based**	**NA**
1. Afzal (2015)	✔	✔		✔		
2. Ai et al. (2015)		✔	✔			
3. Fattoh et al. (2015)	✔	✔	✔		✔	
4. Huang and Mostow (2015); Mostow et al. (2017)	✔					✔
5. Kumar et al. (2015b)	✔	✔			✔	
6. Kumar et al. (2015a)	✔	✔			✔	
7. Lin et al. (2015)		✔	✔			
8. Lopetegui et al. (2015)		✔	NC	NC	NC	NC
9. Majumder and Saha (2015)	✔	✔		✔		
10. Mazidi and Nielsen (2015)	✔	✔	✔			
11. Niraula and Rus (2015)	✔	✔				✔
12. Odilinye et al. (2015)	✔	✔	✔			
13. Polozov et al. (2015)		✔	✔			
14. Shirude et al. (2015)	✔	✔	✔			
15. Singhal et al. (2015a,b, 2016)		✔	NC	NC	NC	NC
16. Susanti et al. (2015, 2016, 2017a,b)	✔	✔	✔			
17. Tamura et al. (2015)	✔	✔		✔		
18. Vinu and Kumar (2015b)		✔	✔			
19. Vinu and Kumar (2015a, 2017a,b); Vinu et al. (2016)		✔	✔			
20. Zhang and Takuma (2015)	✔					✔
21. Alsubait et al. (2016); Kurdi et al. (2017)		✔	✔			
22. Araki et al. (2016)		✔	✔			
23. Hill and Simha (2016)	✔	✔				✔
24. Huang and He (2016)	✔	✔		✔		
25. Jouault et al. (2015a,b, 2016a,b, 2017)		✔	✔			
26. Kwankajornkiet et al. (2016)	✔	✔				✔
27. Mazidi and Tarau (2016a,b)	✔	✔	✔			
28. Shenoy et al. (2016)		✔	NC	NC	NC	NC
29. Song and Zhao (2016a,b)		✔	✔			
30. Wang and Su (2016)		✔	✔			
31. Zhang and VanLehn (2016)		✔	✔			
32. Adithya and Singh (2017)	✔	✔		✔		
33. Blšták (2018); Blšták and Rozinajová (2017)	✔	✔		✔		
34. Chinkina et al. (2017)	✔			✔		
35. Chinkina and Meurers (2017)	✔	✔	✔	✔		
36. Das and Majumder (2017)	✔					✔
37. Gupta et al. (2017)		✔	✔			
38. Jiang and Lee (2017)	✔	✔				✔
39. Kaur and Singh (2017)	✔			✔		
40. Liang et al. (2017)	NC	NC				✔
41. Liu et al. (2017)	✔	✔		✔		
42. Olney et al. (2017)	✔					✔
43. Santhanavijayan et al. (2017)	✔	✔		✔		
44. Satria and Tokunaga (2017a,b)	✔		✔			
45. Seyler et al. (2017)		✔	✔			
46. Shah et al. (2017)	✔	✔				✔
47. Soonklang and Muangon (2017)	✔					✔
48. Stasaski and Hearst (2017)		✔		✔		
49. Basuki and Kusuma (2018)	✔	✔	✔			
50. Blšták and Rozinajová (2018)	✔	✔			✔	

Table 15 (continued)

51. Faizan and Lohmann (2018); Faizan et al. (2017)		✔	✔			
52. Flor and Riordan (2018)	✔	✔		✔		
53. Gao et al. (2018)		✔			✔	
54. Killawala et al. (2018)	✔	✔		✔	✔	
55. Khodeir et al. (2018)		✔	NC	NC	NC	NC
56. Kumar et al. (2018)	✔	✔			✔	
57. Kusuma and Alhamri (2018)	✔		NC	NC	NC	NC
58. Lee et al. (2018)	✔	✔		✔		
59. Liang et al. (2018)	✔	✔				✔
60. Liu et al. (2018)	✔	✔				✔
61. Marrese-Taylor et al. (2018)	✔				✔	
62. Mazidi (2018)	✔	✔	✔			
63. Park et al. (2018)	✔	✔				✔
64. Patra and Saha (2018a,b)		✔				✔
65. Rocha and Zucker (2018)		✔	NC	NC	NC	NC
66. Wang et al. (2018)	✔	✔			✔	
67. Wita et al. (2018)		✔				✔
68. Yaneva et al. (2018)		✔				✔
69. Zhang et al. (2018)	✔	✔		✔		
70. Zavala and Mendoza (2018)		✔	✔			
71. Kurdi et al. (2019); Leo et al. (2019)		✔	✔			
72. Thomas et al. (2019)	✔		✔			
Total	45	60	27	16	9	17

Evaluation

Table 16 Evaluation metrics and results. Studies with multiple evaluations that report on the same metric are in separate rows. Note that in cases where multiple methods are proposed and evaluated in the same study, we report on the results of the best performing method. #Q = no of evaluated questions; #P = no of participants (whether S = student(s), E = expert(s), C = co-worker(s) or A = author(s)); avg. = average; NR = not reported in the paper; NC = not clear; NA = not applicable; * = not reported but calculated based on provided data; and (+) = refer to the paper for extra information about the results or the context of the study

Reference	Domain	#Q	#P	Result	Metric
Statistical difficulty					
Susanti et al. (2015, 2016, 2017a, b)	language	50	79 S	MQs: ranged from 0.18 to 0.90 (mean = 0.51, SD = 0.2) (+)	NA
Huang and He (2016)	language (RC)	24	42 S	avg. of 0.54	
Liu et al. (2018)	language	25	296 S	avg. of 0.68 (+)	
Satria and Tokunaga (2017a, b)	language	30	81 S	ranged from .20 to .96 (mean = .59, SD = .24) (+)	
Lin et al. (2015)	generic	45	30 C	NC	
Vinu and Kumar (2015a, 2017a, b); Vinu et al. (2016)	generic	24	54 S	70.83% agreement between actual and predicted difficulty (+)	
Alsubait et al. (2016) and Kurdi et al. (2017)	generic	12	26 S	7 questions in line with difficulty prediction	
Shenoy et al. (2016)	computer science	4	23 S	NR	
Shenoy et al. (2016)	computer science	4	23 S	NR	time taken to answer
Polozov et al. (2015)	math	25	1,000 C	73% of auto-generated questions answered correctly	NA
Wang and Su (2016)	math	24	30 S	difficulty of auto-generated questions is similar to human-authored questions (+)	

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Difficulty					
Ai et al. (2015)	language (RC)	NR	5 S	NR	5-point scale (categories NR)
Gao et al. (2018)	language (RC)	200	5 NC	1.11 for easy and 1.21 for hard questions	3-point scale (3: top difficulty)
Wita et al. (2018)	language	240	4 E	35% easy (corresponds to very easy + easy); 59.17% moderate (corresponds to reasonable difficult), 5.83% difficult (corresponds to quite difficult)	4-point scale (1: very easy; 2: reasonable easy; 3: reasonable difficult; 4: quite difficult)
Vinu and Kumar (2015b)	generic	31 (75 option-sets)	7 E	65.33% cases agreement with reviewers	3-point scale (low; medium; high)
Adithya and Singh (2017)	generic	NR	NR	40% easy	NR
	generic	14	50 E	84.7% reviewers agreed on easy questions and 38.5% reviewers agreed on difficult questions	
Khodeir et al. (2018)	math	25	4 E	NC	9-point scale (from 1: extremely easy to 9: extremely difficult)
Thomas et al. (2019)	computer science	36	12 E	230 (out of 430) difficulty labels assigned by experts were in line with tool assigned difficulty	3-point scale
Question acceptability or overall quality					
Afzal (2015)	generic	80	2 E	avg. of 2.24	5-point scale (from 0: unacceptable to 5: acceptable)
Mazidi and Nielsen (2015)	generic	NR	NR	avg. of 2.65	3-point scale (1: not acceptable; 2: borderline acceptable; 3: acceptable)
Mazidi and Tarau (2016a, b)	generic	200	NR	72% acceptable questions	5-point scale (categories NR)

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Adithya and Singh (2017)	generic	NR	NR	80% valid questions (+)	NR
Santhanavijayan et al. (2017)	generic	93	NR	67 usable questions	NC
Shah et al. (2017)	generic	NR	NR	avg. of 70.66% acceptable questions (across paragraphs)	binary scale (acceptable; not acceptable)
Stasaski and Hearst (2017)	generic	90	3 E	avg. of 4.15 and 3.89 for for two relation questions and three relation questions respectively	7-point scale (1: poor; 4: OK; 7: excellent)
Basuki and Kusuma (2018)	generic	386	NR	314 true and 42 understandable (+)	3-point scale (true; understandable; false)
Blšták and Rozinajová (2018)	generic	2,564	NR	2,296 acceptable	3-point scale
Huang and Mostow (2015, 2017)	language (RC)	13	8 E	avg. of 2.04	3-point scale (bad; ok; good)
Susanti et al. (2015, 2016, 2017a, b)	language	75	7 E	59% received average score more than or equals to 3	5-point scale
Huang and He (2016)	language	NR	4 E	56.2% acceptable questions	binary scale (acceptable; deficient)
Kwankajornkiet et al. (2016)	language (RC)	394	3 E	73.86% acceptable questions	binary scale (1: acceptable; 0: not acceptable)
Liu et al. (2017)	language (RC)	600	3 E	79% (for top 50)	binary scale
Lee et al. (2018)	language (RC)	2,693	NR	50% acceptable	
Singhal et al. (2015a, b, 2016)	multiple	NR	10 E	80% reviewers are very satisfied with quality	5-point scale (very satisfied; satisfied; neutral; unsatisfied; very unsatisfied)
Thomas et al. (2019)	computer science	200	NR	ranges between 53.6% to 93% acceptable	NR

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Grammatical correctness					
Araki et al. (2016)	language (RC)	200	2 E	avg. of 1.5	3-point scale (1: no grammatical errors; 2 :1 or 2 grammatical errors; 3: 3 or more grammatical errors)
Blšták and Rozinajová (2017) and Blšták (2018)	language (RC)	2,564	NR	89.55% correct questions*	binary scale (correct; not correct)
Blšták and Rozinajová (2017) and Blšták (2018)	language (RC)	122	NR	80.33% correct questions*	binary scale (correct; not correct)
Blšták and Rozinajová (2017) and Blšták (2018)	language (RC)	100	NR	.74	3-point scale (1: correct; 0.5: almost correct; -1: incorrect)
Chinkina and Meurers (2017)	language	69	364 C	avg. of 4.40	5-point scale (categories NR)
Chinkina et al. (2017)	language	69	364 C	NR	5-point scale (categories NR)
Chinkina et al. (2017)	language	96	477 C	NR	5-point scale (categories NR)
Kumar et al. (2018)	language (RC)	700	3 E	63% correct	binary scale (correct; not correct)
Alsubait et al. (2016) and Kurdi et al. (2017)	generic	506	1 E	72.53% questions require minor correction	3-point scale (minor correction; medium correction; major correction)
Mazidi (2018)	generic	149	NR	avg. of 4.1	5-point scale
Blšták and Rozinajová (2018)	generic	100	NR	0.74	3-point scale
Faizan and Lohmann (2018) and Faizan et al. (2017)	generic	14	50 E	28% rated 5 and 40% rated 4	5-point scale (categories NR)
Flor and Riordan (2018)	generic	890	2 E	avg. of 4.32 and 3.89 for yes/no questions and other questions respectively	5-point scale (5: grammatically well-formed; 4: mostly well-formed, with slight problems; 3: has grammatical problems; 2: seriously disfluent; 1: severely mangled)

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Patra and Saha (2018a) and Patra and Saha (2018b)	sport	200	5 E	2.88	3-point scale (0: not acceptable; 0.5: maybe used but better distractors are there; 1: perfect)
Kaur and Singh (2017)	NC	1,220	NR	77.8% grammatically correct questions	NR
Zhang et al. (2018)	medicine	600	NR	517.6 grammatically well-formed	NR
Tamura et al. (2015)	history	100	3 NC	86.5% correct	NR
Semantic ambiguity					
Afzal (2015)	generic	80	2 E	avg. of 2.63	3-point scale (from 1: incomprehensible to 3: clear)
Blšták and Rozinajová (2018)	generic	100	NR	0.68	3-point scale
Flor and Riordan (2018)	generic	890	2 E	avg. of 4.34 and 3.79 for Y/N and other question respectively	5-point scale (5: semantically adequate; 4: mostly semantically adequate, with slight problems; 3: has semantic problems; 2: serious misunderstanding of the original sentence; 1: severely mangled and makes no sense)
Kusuma and Alhamri (2018)	generic	654	1 E	534 questions were declared to be valid and 120 questions declared invalid	NR
Blšták and Rozinajová (2017, 2018)	language (RC)	100	NR	.68	3-point scale (1: correct; -1: incorrect, 0.5: almost correct; 0: not sure)
Kumar et al. (2018)	language (RC)	700	3 E	61% semantically correct	binary scale (correct; not correct)

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Polozov et al. (2015)	math	25	1,000 C	NC	NR
Khodeir et al. (2018)	math	25	4 E	avg. of 3.53	4-point scale (1: very unsatisfactory and unclear, 4: very satisfactory and clear)
Zhang and VanLehn (2016)	biology	40	12 S	3.97 (SD = 0.33)	5-point scale (1: not at all, 5: very ambiguous)
Jouault et al. (2015a, b, 2016a, b, 2017)	history	NR	12 S	2.42 (SD = 1.51)	5-point scale (categories NR)
Zhang et al. (2018)	medicine	600	NR	554.4 semantically adequate	NR
Educational usefulness					
Vinu and Kumar (2015b)	generic	31	7 E	70.97%* useful	3 categories (useful; not useful, but domain related ; not useful and not domain related)
Alsubait et al. (2016) and Kurdi et al. (2017)	generic	115	3 E	94.78% useful by at least one reviewer	
Flor and Riordan (2018)	generic	890	2 E	71% and 50% useful questions for Y/N questions and other questions respectively	NA
Jouault et al. (2015a, b, 2016a, b, 2017)	history	60	1 E	76.67%* questions rated 3 or more	5-point scale (5: questions contribute to deepening the understanding of the learners; 1: do not)
Tamura et al. (2015)	history	100	3 NC	48% appropriate	scale NR
Satria and Tokunaga, (2017a, b)	language	60	5 E	65% acceptable questions	3-point scale (1: problematic; 2: acceptable but can be improved; 3: acceptable)

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Gupta et al. (2017)	math	8	12 S	avg. of 3.75 (SD = 0.62)	5-point scale (categories NR)
Singhal et al. (2015a, b, 2016)	multiple domains	NR	10 E	80% and 100% of the reviewers indicate that they will use the generator for assessment and teaching respectively	6-point scale (definitely; probably; neutral; probably not; definitely not; not applicable)
Zhang and VanLehn (2016)	biology	40	12 S	2.72 (SD = 0.34)	5-point scale (from 5: yes to 1: not at all)
Leo et al. (2019)	medicine	435	15 E	79% appropriate by at least one reviewer	binary scale (appropriate, inappropriate)
Relevance to the input					
Kumar et al. (2015b)	generic	495	15 S	241 questions rated 3, 164 rated 2, 59 rated 1 and 31 rated 0	4-point scale (0: Sentence is bad; Does not matter whether gap and distractors are good or bad; 1: Sentence is good, gap is bad; Does not matter whether distractors are good or bad; 2: Sentence and gap are good but distractors are bad; 3: Sentence, gap and distractors are all good)
Odilinye et al. (2015)	generic	NC	1 NC	range: 40 to 100	NA
Mazidi (2018)	generic	149	NR	avg. of 4.3	5-point scale
Faizan and Lohmann (2018) and Faizan et al. (2017)	generic	14	50 E	40.33% gap-fill questions had the highest relevance, while 8.67% and 10% chose the type and Jeopardy questions had the highest relevance (+)	5-point scale (categories NR)
Flor and Riordan (2018)	generic	890	2 E	avg. of 2.75 and 2.52 for Y/N and other questions respectively	4-point scale (3: is about the sentence; 2: goes beyond the information in the sentence; 1: veers away, is unrelated to the sentence; 0: too mangled to make a reasonable judgment)

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Wang et al. (2018)	generic	300	NR	> 80% relevant	binary scale (Yes; No)
Soonklang and Muangon (2017)	language	NR	NR	avg. of 99.05	NR
Kumar et al. (2018)	language (RC)	700	3 E	67% relevant	binary scale (relevant; not relevant)
Domain Relevance					
Afzal (2015)	generic	80	2 E	avg. of 1.85	3-point scale (from 1: not relevant to 3: very relevant)
Alsubait et al. (2016) and Kurdi et al. (2017)	generic	65	3 E	100% relevant	binary choice
Song and Zhao (2016a, b)	generic	1,000	3E	F-score = 44.6 and 34.1 for development and test sets respectively	3-point scale (categories NR)
Rocha and Zucker (2018)	generic	200	1 E	avg. of 3.25 for KB1; 2.87 for KB2	5-point scale (5: the most relevant)
Zhang and VanLehn (2016)	biology	40	12 S	3.51 (SD = 0.46)	5-point scale (from 1: not at all to 5: all of them)
Gao et al. (2018)	language (RC)	200	5 NC	0.75 and 0.64 relevant for easy for difficult questions respectively	binary scale
Fluency					
Song and Zhao (2016a, b)	generic	1,000	3E	F-score = 93.2 and 94.5 for development and test sets respectively	3-point scale (categories NR)
Wang et al. (2018)	generic	300	NR	> 70% fluent	binary rating (Yes; No)
Zhang and VanLehn (2016)	biology	40	12 S	4.05 (SD = 0.34)	5-point scale (from 5: very natural to 1: not at all)
Gao et al. (2018)	language (RC)	200	5 NC	2.93 for easy 2.89 for difficult questions	3-point scale (3: top fluency)

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Khodeir et al. (2018)	math	25	4 E	avg. of 3.2	4-point scale (1: very unsatisfactory and unclear and 4: very satisfactory and clear)
Jouault et al. (2015a, b, 2016a, b, 2017)	history	NR	12 S	2.67 (SD = 1.44)	5-point scale (categories NR)
Being indistinguishable from human-authored questions					
Wang and Su (2016)	math	24	30 S	No significant difference between the two groups	
Khodeir et al. (2018, 2018)	math	25	4 E	82% questions were thought to be human-authored	3 categories (system-generated; human-generated; unsure)
Susanti et al. (2015, 2016, 2017a, b)	language	22	7 E	45% questions were thought to be human-authored	binary choice (human-generate; machine-generated)
Chinkina and Meurers (2017)	language	69	364 C	67% questions were thought to be human-authored	binary choice
Killawala et al. (2018)	generic	NR	NR	NC	5-point scale
Wang et al. (2018)	generic	300	NR	> 60% questions were thought to be human-authored	binary choice
Overlap with human generated questions					
Shirude et al. (2015)	generic	NR	12 E	63.15% types of questions generated by the human are covered by the generator	NR
Vinu and Kumar (2015a, 2017a, b, 2016)	generic	NR	NR	recall = 43% to 81%, precision = 72% to 93% (+)	
Liu et al. (2017)	language (RC)	600	3 E	recall = 64%, precision = 69% (+)	NR
Jouault et al. (2015a, b, 2016a, b, 2017)	history	69	1 E	84% of the human-authored questions are covered by auto-generated questions	coverages means that both questions cover the same knowledge

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Kaur and Singh (2017)	NC	1,220	NR	recall = 73.93%	NR
Discrimination					
Susanti et al. (2015, 2016, 2017a, b)	language	50	79 S	74% questions have acceptable discrimination (≥ 0.2)	NA
Huang and He (2016)	language (RC)	24	42 S	avg. of 0.36	
Liu et al. (2018)	language	25	296 S	avg. of 0.41 (+)	
Satria and Tokunaga (2017a, b)	language	30	81 S	73.33% questions have acceptable discrimination (≥ 0.2) (mean = 0.33) (+)	
Alsubait et al. (2016) and Kurdi et al. (2017)	generic	12	26 S	discrimination was greater than 0.4. for 10 questions	
ROUGE					
Odilinye et al. (2015)	generic	NC	1 S	range: between 15 and 30 (on their dataset)	
Blšták and Rozinajová (2018)	generic	66	–	0.86 (on QGSTEC)	
Wang et al. (2018)	generic	NR	NA	44.37 (on SQUAD)	
Blšták and Rozinajová (2017) and Blšták (2018)	language (RC)	NR	NA	83	
Gao et al. (2018)	language (RC)	NC	NA	**46.22** (on SQUAD)	
Kumar et al. (2018)	language (RC)	700	NA	41.75 (on SQUAD)	
BLEU					
Blšták and Rozinajová (2018)	generic	66	–	B1 = 79, B2 = 75, B3 = 72 and B4 = 70 (on QGSTEC)	
Wang et al. (2018)	generic	NR	NA	B4 = 13.86 (on SQUAD)	
Blšták and Rozinajová (2017) and Blšták (2018)	language (RC)	NR	NA	75	
Gao et al. (2018)	language (RC)	NC	NA	B1 = 44.11, B2 = **29.64**, B3 = **21.89** and B4 = **16.68** (on SQUAD)	
Kumar et al. (2018)	language (RC)	700	NA	B1 = **46.32**, B2 = 28.81, B3 = 19.67 and B4=13.85 (on SQUAD)	

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Freeness from error					
Huang and He (2016)	language	24	1 S	25% received no modification and 33% minor modification	3-point scale (no modification; involve insertion, deletion and reordering in no more than two positions, and maintained the basic grammatical structures; involved modification beyond those defined in the previous group)
Satria and Tokunaga (2017a, b)	language	100	1 A	53% error free	NA
Afzal (2015)	generic	80	2 E	avg. of 2.36	4-point scale (1: unusable; 2: minor revisions; 3: major revisions; 4: directly usable)
Alsubait et al. (2016) and Kurdi et al. (2017)	generic	506	1 E	4.9% flawless question	3-point scale (flawless; 1 Flaw; $\geq$ 2 Flaws)
METEOR					
Kumar et al. (2018)	language (RC)	700	NA	18.51 (on SQUAD)	
Gao et al. (2018)	language (RC)	NC	NA	**20.94**	
Wang et al. (2018)	generic	NR	NA	18.38	
Answerability					
Araki et al. (2016)	language (RC)	200	2 E	avg. of 1.21	binary-scale (1: yes, 2: no)
Chinkina and Meurers (2017)	language	69	364 C	avg. of 4.47	5-point scale (categories NR)
Chinkina et al. (2017)	language	69	364 C	NR	5-point scale (categories NR)
Chinkina et al. (2017)	language	96	477 C	NR	5-point scale (categories NR)
Cognitive level or depth					
Araki et al. (2016)	language (RC)	200	2 E	avg. of 0.76	Number of inference steps
Zhang and VanLehn (2016)	biology	40	12 S	2.59 (SD = 0.35)	5-point scale (1: Just memorizing to 5: thinking)

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Learning outcome					
Olney et al. (2017)	language (RC)	53 and 32	302 C	generated questions had significantly higher post-test proportion correct that other type of questions (+)	
Zavala and Mendoza (2018)	computer science	4 sets (12 each)	17	students showed an improvement in their skills (+)	
Diversity of question types					
Huang and He (2016)	language (RC)	459	NA	The majority are what questions (232 questions) followed by who questions (109 questions) (+)	
Soonklang and Muangon (2017)	language	NG	NA	NG	
How much the questions revealed about the answer					
Faizan and Lohmann (2018) and Faizan et al. (2017)	generic	14	50 E	The majority of questions rated as 3 or 4 (+)	5-point scale (categories NR)
Distractor quality or plausibility					
Afzal (2015)	generic	80	2 E	avg. of 3.16	5-point scale (from 0: unacceptable to 5: acceptable)
Kumar et al. (2015a)	generic	75	NR	43% and 51% of the distractors being very good and fair respectively	binary rating (1: good; 0: bad)
Liang et al. (2017)	generic	122	3 E	51.7% and 48.4% acceptable distractors (good and fair) for Wiki-FITB and Course-FITB receptively	3-point scale (good; fair; bad)
Santhanavijayan et al. (2017)	generic	NC	NR	41 questions have distractors that are very close to the key	NM
Seyler et al. (2017)	generic	400	NR	76% agreement between reviewer and generator, Cohen's kappa of 0.52	NR

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Shah et al. (2017)	generic	NR	NR	avg. of 61.71% acceptable distractors (across paragraphs)	binary scale (acceptable; not acceptable)
Stasaski and Hearst (2017)	generic	20	1 E	2.78	5-point scale (categories NG)
Faizan and Lohmann (2018) and Faizan et al. (2017)	generic	14	50 E	avg. of 2 and 3 for easy and hard distractors respectively	5-point scale (categories NR)
Park et al. (2018)	generic	50	10 S	avg of 2.08	number of good distractors
Araki et al. (2016)	language (RC)	200	2 E	avg. of 1.94	3-point scale (1: A distractor is confusing because it overlaps the correct answer partially or completely; 2: A distractor can be easily identified as an incorrect answer; 3: A distractor can be viable)
Hill and Simha (2016)	language (RC)	30	67 NC	avg. of 58% distractors fit blanks in a narrow context	
Kwankajornkiet et al. (2016)	language (RC)	394	3 E	27.25% acceptable distractors	binary scale (1: acceptable; 0: not acceptable)
Jiang and Lee (2017)	language	37	2 E	46.6%	3-point scale 3: plausible; 2: Somewhat plausible; 1: obviously wrong)
Majumder and Saha (2015)	sport	112	5 NC	91.07% accuracy in distractor selection	NR
Patra and Saha (2018a, 2018b)	sport	200	5 E	2.71	3-point scale (1: perfect; 0.5: maybe used but better distractors are there; 0: not acceptable)
Patra and Saha (2018a, 2018b)	sport	100	3 E	avg of 87.7%	binary scale

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Patra and Saha (2018a, b)	sport	100	3 E	avg of 79.3%	binary scale
Patra and Saha (2018a, b)	sport	200	15 NC	avg of 2.3	3-point scale (3: all the distractors are good; 0: none of distractors is good)
Leo et al. (2019)	medicine	435	15 E		4-point scale (not plausible, plausible but easy to eliminate, difficult to eliminate, cannot eliminate)
Answer correctness					
Araki et al. (2016)	language (RC)	200	2 E	avg. of 1.46	3-point scale (1: correct; 2: partially correct; 3: incorrect)
Hill and Simha (2016)	language (RC)	30	67 NC	94.6% of keys fits blank in broad context	NA
Majumder and Saha (2015)	sport	112	5 NC	83.03% accuracy in key selection	NR
Distractor correctness					
Jiang and Lee (2017)	language	37	2 E	93.2% to 100% incorrect	binary scale (correct; incorrect)
Distractor functionality					
Alsubait et al. (2016) and Kurdi et al. (2017)	generic	6	19 S	at least two out of three distractors were useful	
Alsubait et al. (2016) and Kurdi et al. (2017)	generic	6	7 S	at least one distractor were useful except for one question	
Liu et al. (2018)	language	75 distractors	296 S	76% useful distractors*	
Overlap with human-generated distractors					
Liang et al. (2018)	generic	20.8K	NA	a recall of about 90% (+)	
Yaneva and et al. (2018)	medicine	1810	NA	1 in 5 distractors match human-distractors when producing 20 candidates for each item	

Table 16 (continued)

Reference	Domain	#Q	#P	Result	Metric
Distractor homogeneity					
Faizan and Lohmann (2018) and Faizan et al. (2017)	generic	14	50 E	NR	
Option usefulness					
Wita et al. (2018)	language	240	2 E & 2 S	1.67% ambiguous; 10% less-related choices; 88.33% more related choices	4 categories (1: not useful because redundant choices; 2: not useful because ambiguous; 3: useful with broad relationship between choices; 4: useful with close relationship between choices)
Distractor matching intended type					
Mostow et al. (2017) and Huang and Mostow (2015)	language (RC)	16	5 E	Cohen's Kappa for reviewer agreement with type ranged from 0.48 to 0.81	
Distractor readability					
Afzal (2015)	generic	80	2 E	avg. of 2.63	3-point scale (from 1: incomprehensible to 3: clear)
Blank quality					
Santhanavijayan et al. (2017)	generic	NC	NR	52 questions with a blank at a suitable positing	NR
Park et al. (2018)	generic	50	10 S	avg. of 0.8	binary scale (0: bad; 1: good)
Marrese-Taylor et al. (2018)	language	1.5M	NA	89.31% accuracy of blanked words	
Generality of the designed templates					
Odilinye et al. (2015)	generic	NC	1 NC	38.2 to 29.4% of templates are used	NA
Sentence quality					
Park et al. (2018)	generic	50	10 S	avg. of 0.75	binary scale (0: bad; 1: good)

Quality assessment

Table 17 Quality assessment of reviewed literature (✔ = yes; ✗ = no; NS = not specified; NC = not clear; and NA = not applicable)

Reference	Q1	Q2	Q3	Q4	Q5	Q6a	Q6b	Q7	Q8	Q9
1. Afzal (2015)	✔	✔	✗	✗	✔	✗	✗	NS	✗	✔
2. Ai et al. (2015)	✔	✔	✗	NS	✗	✗	✗	NS	✔	✗
3. Fattoh et al. (2015)	NA	NA	NA	NA	NC	✗	✗	NS	✔	NA
4. Huang and Mostow (2015); Mostow et al. (2017)	✔	✔	✔	✔	✔	✗	✗	NS	✔	NA
5. Kumar et al. (2015b)	✔	✔	✗	✗	✔	✗	✗	NS	✔	✗
6. Kumar et al. (2015a)	✗	✗	✔	✗	✔	✗	✗	NS	✔	✔
7. Lin et al. (2015)	✔	✗	✗	NS	✔	✔	✗	NS	✔	NA
8. Lopetegui et al. (2015)	✔	✔	✗	✗	NC	✔	✗	✔	✔	✗
9. Majumder and Saha (2015)	✔	✗	✗	NS	✔	✗	✗	NS	✔	✗
10. Mazidi and Nielsen (2015)	✗	✗	✔	NS	✗	✔	NA	NS	✔	✔
11. Niraula and Rus (2015)	NA	NA	NA	NA	✔	✔	✔	✔	✔	NA
12. Odilinye et al. (2015)	✗	✔	✗	NS	✔	✔	✗	NS	✔	✗
13. Polozov et al. (2015)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✗
14. Shirude et al. (2015)	✔	✗	✗	NS	✗	✗	✗	NS	✔	✗
15. Singhal et al. (2015a,b, 2016)	✔	✔	✗	✗	✗	✗	✗	NS	✔	✗
16. Susanti et al. (2015, 2016, 2017a,b)	✔	✔	✗	✗	✔	✗	✗	NS	✔	✗
17. Tamura et al. (2015)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✗
18. Vinu and Kumar (2015a, 2017a,b); Vinu et al. (2016)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✗
19. Vinu and Kumar (2015b)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✗
20. Zhang and Takuma (2015)*	✔	✗	✗	NS	✗	✗	✗	NS	✔	✗
21. Alsubait et al. (2016); Kurdi et al. (2017)	✔	✔	✗	✗	✔	✔	✔	✔	✔	✔
22. Araki et al. (2016)	✔	✗	✗	NS	✔	✗	✗	NS	✔	✔
23. Hill and Simha (2016)	✔	✔	✔	✔	✔	✔	✗	NS	✔	✗
24. Huang and He (2016)	✔	✔	✗	NS	✔	✔	✗	NS	✔	✗
25. Jouault et al. (2015a,b, 2016a,b, 2017)	✔	✔	✗	✗	✔	✗	✗	NS	✔	✗
26. Kwankajornkiet et al. (2016)	✔	✔	✗	NS	✔	✔	NA	NS	✔	✗
27. Mazidi and Tarau (2016a,b)	✗	✔	✔	✔	✔	✔	✗	NS	✔	✔
28. Shenoy et al. (2016)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✗

Table 17 (continued)

29. Song and Zhao (2016a,b)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✔
30. Wang and Su (2016)*	✔	✔	✗	✗	✔	✗	✗	NS	✔	NA
31. Zhang and VanLehn (2016)*	✔	✔	✔	✔	✔	✔	✗	NS	✔	✗
32. Adithya and Singh (2017)	✗	✗	✗	NS	✗	✗	✗	NS	✗	✗
33. Blšták (2018); Blšták and Rozinajová (2017)	✔	✗	✗	NS	✔	NC	✗	NS	✔	✗
34. Chinkina and Meurers (2017)	✔	✔	✔	✔	✔	NC	NC	NS	✔	✗
35. Chinkina et al. (2017)	✔	✔	✔	✔	✔	✗	✗	NS	✗	✔
36. Das and Majumder (2017)	✔	✔	✗	NS	✔	✗	✗	NS	✔	✔
37. Gupta et al. (2017)	✔	✔	✔	✔	✔	✔	✗	NS	✔	NA
38. Jiang and Lee (2017)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✔
39. Kaur and Singh (2017)	✗	✗	✗	NS	✔	✔	NA	✔	✗	✗
40. Liang et al. (2017)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✗
41. Liu et al. (2018)	✔	✔	✔	✔	✔	✔	NA	✔	✔	NA
42. Olney et al. (2017)	✔	✔	✔	✔	✔	✔	✗	NS	✔	NA
43. Santhanavijayan et al. (2017)	✗	✗	✗	NC	✔	✗	✗	NC	✔	✗
44. Satria and Tokunaga (2017a,b)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✔
45. Seyler et al. (2017)	✔	✗	✗	NS	✔	✔	✗	NS	✔	✔
46. Shah et al. (2017)	✗	✔	✗	✗	✗	✔	✗	NS	✔	✗
47. Soonklang and Muangon (2017)	✔	✔	✗	NS	✗	✗	✗	NS	✔	✗
48. Stasaski and Hearst (2017)	✔	✔	✔	NS	✔	✔	✗	NS	✔	✗
49. Basuki and Kusuma (2018)	✗	✗	✗	NS	✔	✗	✗	NS	✔	✗
50. Blšták and Rozinajová (2018)	✗	✗	✗	NS	✔	✗	✗	NS	✔	✗
51. Faizan and Lohmann (2018); Faizan et al. (2017)	✔	✗	✔	NS	✔	✔	✔	NS	✔	✗
52. Flor and Riordan (2018)	✔	✔	✗	NS	✔	✗	✗	NS	✔	✔
53. Gao et al. (2018)	✔	✗	✗	NS	✔	✔	✗	NS	✔	✗
54. Killawala et al. (2018)	✗	✗	✗	NS	✗	✗	✗	NS	✗	✗
55. Khodeir et al. (2018)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✔
56. Kumar et al. (2018)	✔	✔	✗	NS	✔	✔	✗	NS	✔	✗
57. Kusuma and Alhamri (2018)	✔	✗	✗	NS	✔	✔	NA	NS	✔	✗
58. Lee et al. (2018)	✗	✗	✗	NS	✔	✗	✗	NS	✗	✗
59. Liang et al. (2018)	NA	NA	NA	NA	✔	✔	✗	✔	✗	NA

Table 17 (continued)

60. Liu et al. (2017)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✔
61. Marrese-Taylor et al. (2018)	NA	NA	NA	NA	✔	✔	✗	NS	✗	NA
62. Mazidi (2018)	✗	✔	✔	✗	✔	✔	✗	✔	NS	✗
63. Park et al. (2018)	✔	✗	✗	NS	✔	✔	✗	NS	✔	✗
64. Patra and Saha (2018a,b)	✔	✔	✗	NS	✔	✗	✗	NS	✔	✗
65. Rocha and Zucker (2018)	✔	✔	✗	✗	✔	✔	✗	NS	✔	✗
66. Wang et al. (2018)	✗	✗	✗	NS	✗	✗	✗	NS	✔	✗
67. Wita et al. (2018)	✔	✔	✗	✗	✔	✗	✗	NS	✔	✗
68. Yaneva et al. (2018)	NA	NA	NA	NA	✔	✔	NA	✔	✔	NA
69. Zhang et al. (2018)	✗	✗	✗	NS	✔	✗	✗	NS	✗	✗
70. Zavala and Mendoza (2018)	✔	✔	✗	✗	✔	✗	✗	NS	✗	NA
71. Leo et al. (2019)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
72. Thomas et al. (2019)	✔	✔	✔	NS	✔	✔	✔	✔	✔	✔

References

Abacha, A.B., & Demner-Fushman, D. (2016). Recognizing question entailment for medical question answering. In: the AMIA annual symposium, American medical informatics association, p. 310.

Adithya, S.SR., & Singh, P.K. (2017). Web authoriser tool to build assessments using Wikipedia articles. In: TENCON 2017 - 2017 IEEE region 10 conference, pp. 467–470. https://doi.org/10.1109/TENCON.2017.8227909.

Afzal, N. (2015). Automatic generation of multiple choice questions using surface-based semantic relations. *International Journal of Computational Linguistics (IJCL)*, *6*(3), 26–44. https://doi.org/10.1007/s00500-013-1141-4.

Afzal, N., & Mitkov, R. (2014). Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Computing*, *18*(7), 1269–1281. https://doi.org/10.1007/s00500-013-1141-4.

Afzal, N., Mitkov, R., Farzindar, A. (2011). Unsupervised relation extraction using dependency trees for automatic generation of multiple-choice questions. In: Canadian conference on artificial intelligence, Springer, pp. 32–43. https://doi.org/10.1007/978-3-642-21043-3_4.

Ai, R., Krause, S., Kasper, W., Xu, F., Uszkoreit, H. (2015). Semi-automatic generation of multiple-choice tests from mentions of semantic relations. In: the 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pp. 26–33.

Alsubait, T. (2015). *Ontology-based question generation*. PhD thesis: University of Manchester.

Alsubait, T., Parsia, B., Sattler, U. (2012a). Automatic generation of analogy questions for student assessment: an ontology-based approach. Research in Learning Technology 20. https://doi.org/10.3402/rlt.v20i0.19198.

Alsubait, T., Parsia, B., Sattler, U. (2012b). Mining ontologies for analogy questions: A similarity-based approach. In: OWLED.

Alsubait, T., Parsia, B., Sattler, U. (2012c). Next generation of e-assessment: automatic generation of questions. *International Journal of Technology Enhanced Learning*, *4*(3-4), 156–171.

Alsubait, T., Parsia, B., Sattler, U. (2013). A similarity-based theory of controlling MCQ difficulty. In *2013 2Nd international conference on e-learning and e-technologies in education* (pp. 283–288). ICEEE: IEEE. https://doi.org/10.1109/ICeLeTE.2013.664438.

Alsubait, T., Parsia, B., Sattler, U. (2014a). Generating multiple choice questions from ontologies: Lessons learnt. In: OWLED, Citeseer, pp. 73–84.

Alsubait, T., Parsia, B., Sattler, U. (2014b). Generating multiple questions from ontologies: How far can we go? In: the 1st International Workshop on Educational Knowledge Management (EKM 2014), Linköping University Electronic Press, pp. 19–30.

Alsubait, T., Parsia, B., Sattler, U. (2016). Ontology-based multiple choice question generation. *KI - Kü,nstliche Intelligenz*, *30*(2), 183–188. https://doi.org/10.1007/s13218-015-0405-9.

Araki, J., Rajagopal, D., Sankaranarayanan, S., Holm, S., Yamakawa, Y., Mitamura, T. (2016). Generating questions and multiple-choice answers using semantic analysis of texts. In *The 26th international conference on computational linguistics (COLING*, (Vol. 2016 pp. 1125–1136).

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72.

Basuki, S., & Kusuma, S. F. (2018). Automatic question generation for 5w-1h open domain of Indonesian questions by using syntactical template-based features from academic textbooks. *Journal of Theoretical and Applied Information Technology*, *96*(12), 3908–3923.

Baturay, M. H. (2015). An overview of the world of MOOCs. *Procedia - Social and Behavioral Sciences*, *174*, 427–433. https://doi.org/10.1016/j.sbspro.2015.01.685.

Beck, J.E., Mostow, J., Bey, J. (2004). Can automated questions scaffold children's reading comprehension? In: International Conference on Intelligent Tutoring Systems, Springer, pp. 478–490.

Bednarik, L., & Kovacs, L. (2012a). Automated EA-type question generation from annotated texts, IEEE, SACI. https://doi.org/10.1109/SACI.2012.6250000.

Bednarik, L., & Kovacs, L. (2012b). Implementation and assessment of the automatic question generation module, IEEE, CogInfoCom. https://doi.org/10.1109/CogInfoCom.2012.6421938.

Biggs, J. B., & Collis, K.F. (2014). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Cambridge: Academic Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook i: The cognitive domain vol 19*. New York: David McKay Co Inc.

Blšták, M. (2018). Automatic question generation based on sentence structure analysis . *Information Sciences & Technologies: Bulletin of the ACM Slovakia*, *10*(2), 1–5.

Blšták, M., & Rozinajová, V. (2017). Machine learning approach to the process of question generation. In Blšták, M., & Rozinajová, V. (Eds.) *Text, speech, and dialogue* (pp. 102–110). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-64206-2_12.

Blšták, M., & Rozinajová, V. (2018). Building an agent for factual question generation task. In *2018 World symposium on digital intelligence for systems and machines (DISA)* (pp. 143–150). IEEE. https://doi.org/10.1109/DISA.2018.8490637.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, *32*(suppl_1), D267–D270. https://doi.org/10.1093/nar/gkh061.

Boland, A., Cherry, M. G., Dickson, R. (2013). Doing a systematic review: A student's guide. Sage.

Ch, D.R., & Saha, S.K. (2018). Automatic multiple choice question generation from text: A survey. IEEE Transactions on Learning Technologies https://doi.org/10.1109/TLT.2018.2889100, in press.

Chen, C.Y., Liou, H.C., Chang, J.S. (2006). Fast: an automatic generation system for grammar tests. In: the COLING/ACL on interactive presentation sessions, association for computational linguistics, pp. 1–4.

Chinkina, M., & Meurers, D. (2017). Question generation for language learning: From ensuring texts are read to supporting learning. In: the 12th workshop on innovative use of NLP for building educational applications, pp. 334–344.

Chinkina, M., Ruiz, S., Meurers, D. (2017). Automatically generating questions to support the acquisition of particle verbs: evaluating via crowdsourcing. In: CALL in a climate of change: adapting to turbulent global conditions, pp. 73–78.

Critical Appraisal Skills Programme (2018). CASP qualitative checklist. https://casp-uk.net/wp-content/uploads/2018/03/CASP-Qualitative-Checklist-Download.pdf, accessed: 2018-09-07.

Das, B., & Majumder, M. (2017). Factual open cloze question generation for assessment of learner's knowledge. *International Journal of Educational Technology in Higher Education*, *14*(1), 24. https://doi.org/10.1186/s41239-017-0060-3.

Donnelly, K. (2006). SNOMED-CT: The Advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, *121*, 279–290.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology & Community Health*, *52*(6), 377–384.

Fairon, C. (1999). A web-based system for automatic language skill assessment: Evaling. In: Symposium on computer mediated language assessment and evaluation in natural language processing, association for computational linguistics, pp. 62–67.

Faizan, A., & Lohmann, S. (2018). Automatic generation of multiple choice questions from slide content using linked data. In: the 8th International Conference on Web Intelligence, Mining and Semantics.

Faizan, A., Lohmann, S., Modi, V. (2017). Multiple choice question generation for slides. In: Computer Science Conference for University of Bonn Students, pp. 1–6.

Fattoh, I. E., Aboutabl, A. E., Haggag, M. H. (2015). Semantic question generation using artificial immunity. *International Journal of Modern Education and Computer Science*, *7*(1), 1–8.

Flor, M., & Riordan, B. (2018). A semantic role-based approach to open-domain automatic question generation. In: the 13th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 254–263.

Flórez-Vargas, O., Brass, A., Karystianis, G., Bramhall, M., Stevens, R., Cruickshank, S., Nenadic, G. (2016). Bias in the reporting of sex and age in biomedical research on mouse models. eLife 5(e13615).

Gaebel, M., Kupriyanova, V., Morais, R., Colucci, E. (2014). *E-learning in European higher education institutions: Results of a mapping survey conducted in october-December 2013*. Tech. rep.: European University Association.

Gamer, M., Lemon, J., Gamer, M.M., Robinson, A., Kendall's, W. (2019). Package 'irr'. https://cran.r-project.org/web/packages/irr/irr.pdf.

Gao, Y., Wang, J., Bing, L., King, I., Lyu. MR (2018). Difficulty controllable question generation for reading comprehension. Tech. rep.

Goldbach, I.R., & Hamza-Lup, F.G. (2017). Survey on e-learning implementation in Eastern-Europe spotlight on Romania. In: the Ninth International Conference on Mobile, Hybrid, and On-Line Learning.

Gupta, M., Gantayat, N., Sindhgatta, R. (2017). Intelligent math tutor: Problem-based approach to create cognizance. In: the 4th ACM Conference on Learning@ Scale, ACM, pp. 241–244.

Han, S., Olonisakin, T. F., Pribis, J. P., Zupetic, J., Yoon, J. H., Holleran, K. M., Jeong, K., Shaikh, N., Rubio, D. M., Lee, J. S. (2017). A checklist is associated with increased quality of reporting preclinical biomedical research: a systematic review. *PloS One*, *12*(9), e0183591.

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, *73*(2), 94–97. https://doi.org/10.1080/08832329709601623.

Heilman, M. (2011). *Automatic factual question generation from text*. PhD thesis: Carnegie Mellon University.

Heilman, M., & Smith, N.A. (2009). Ranking automatically generated questions as a shared task. In: The 2nd Workshop on Question Generation, pp. 30–37.

Heilman, M., & Smith, N.A. (2010a). Good question! statistical ranking for question generation. In: Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics, association for computational linguistics, pp. 609–617.

Heilman, M., & Smith, N.A. (2010b). Rating computer-generated questions with mechanical turk. In: the NAACL HLT 2010 workshop on creating speech and language data with amazon's mechanical turk, association for computational linguistics, pp. 35–40.

Hill, J., & Simha, R. (2016). Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In: the 11th workshop on innovative use of NLP for building educational applications, pp. 23–30.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *The Journal of the Pakistan Medical Association (JPMA)*, *62*(2), 142–147.

Huang, Y., & He, L. (2016). Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering*, *22*(3), 457–489. https://doi.org/10.1017/S1351324915000455.

Huang, Y. T., & Mostow, J. (2015). Evaluating human and automated generation of distractors for diagnostic multiple-choice cloze questions to assess children's reading comprehension. In Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (Eds.) *Artificial intelligence in education* (pp. 155–164). Cham: Springer International Publishing.

Huang, Y. T., Tseng, Y. M., Sun, Y. S., Chen, M.C. (2014). TEDQuiz: automatic quiz generation for TED talks video clips to assess listening comprehension. In *2014 IEEE 14Th international conference on advanced learning technologies* (pp. 350–354). ICALT: IEEE.

Jiang, S., & Lee, J. (2017). Distractor generation for Chinese fill-in-the-blank items. In: the 12th workshop on innovative use of NLP for building educational applications, pp. 143–148.

Jouault, C., & Seta, K. (2014). Content-dependent question generation for history learning in semantic open learning space. In: The international conference on intelligent tutoring systems, Springer, pp. 300–305.

Jouault, C., Seta, K., Hayashi, Y. (2015a). A method for generating history questions using LOD and its evaluation. *SIG-ALST of The Japanese Society for Artificial Intelligence*, *B5*(1), 28–33.

Jouault, C., Seta, K., Hayashi, Y. (2015b). Quality of LOD based semantically generated questions. In Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (Eds.) *Artificial intelligence in education* (pp. 662–665). Cham: Springer International Publishing.

Jouault, C., Seta, K., Hayashi, Y. (2016a). Content-dependent question generation using LOD for history learning in open learning space. *New Generation Computing*, *34*(4), 367–394. https://doi.org/10.1007/s00354-016-0404-x.

Jouault, C., Seta, K., Yuki, H., et al. (2016b). Can LOD based question generation support work in a learning environment for history learning?. *SIG-ALST*, *5*(03), 37–41.

Jouault, C., Seta, K., Hayashi, Y. (2017). SOLS: An LOD based semantically enhanced open learning space supporting self-directed learning of history. *IEICE Transactions on Information and Systems*, *100*(10), 2556–2566.

Kaur, A., & Singh, S. (2017). Automatic question generation system for Punjabi. In: The international conference on recent innovations in science, Agriculture, Engineering and Management.

Kaur, J., & Bathla, A. K. (2015). A review on automatic question generation system from a given Hindi text. *International Journal of Research in Computer Applications and Robotics (IJRCAR)*, *3*(6), 87–92.

Khodeir, N. A., Elazhary, H., Wanas, N. (2018). Generating story problems via controlled parameters in a web-based intelligent tutoring system. *The International Journal of Information and Learning Technology*, *35*(3), 199–216.

Killawala, A., Khokhlov, I., Reznik, L. (2018). Computational intelligence framework for automatic quiz question generation. In: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. https://doi.org/10.1109/FUZZ-IEEE.2018.8491624.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*. Tech. rep.: Keele University and University of Durham.

Kovacs, L., & Szeman, G. (2013). Complexity-based generation of multi-choice tests in AQG systems, IEEE, CogInfoCom. https://doi.org/10.1109/CogInfoCom.2013.6719278.

Kumar, G., Banchs, R., D'Haro, L.F. (2015a). Revup: Automatic gap-fill question generation from educational texts. In: the 10th workshop on innovative use of NLP for building educational applications, pp. 154–161.

Kumar, G., Banchs, R., D'Haro, L.F. (2015b). Automatic fill-the-blank question generator for student self-assessment. In: IEEE Frontiers in Education Conference (FIE), pp. 1–3. https://doi.org/10.1109/FIE.2015.7344291.

Kumar, V., Boorla, K., Meena, Y., Ramakrishnan, G., Li, Y. F. (2018). Automating reading comprehension by generating question and answer pairs. In Phung, D., Tseng, V.S., Webb, G.I., Ho, B., Ganji, M., Rashidi, L. (Eds.) *Advances in knowledge discovery and data mining* (pp. 335–348). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-93040-4_27.

Kurdi, G., Parsia, B., Sattler, U. (2017). An experimental evaluation of automatically generated multiple choice questions from ontologies. In Dragoni, M., Poveda-Villalón, M., Jimenez-Ruiz, E. (Eds.) *OWL: Experiences And directions – reasoner evaluation* (pp. 24-39). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-54627-8_3.

Kurdi, G., Leo, J., Matentzoglu, N., Parsia, B., Forege, S., Donato, G., Dowling, W. (2019). A comparative study of methods for a priori prediction of MCQ difficulty. the Semantic Web journal, In press.

Kusuma, S. F., & Alhamri, R. Z. (2018). Generating Indonesian question automatically based on Bloom's taxonomy using template based method. *KINETIK: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *3*(2), 145–152.

Kwankajornkiet, C., Suchato, A., Punyabukkana, P. (2016). Automatic multiple-choice question generation from Thai text. In: the 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1–6. https://doi.org/10.1109/JCSSE.2016.7748891.

Le, N. T., Kojiri, T., Pinkwart, N. (2014). Automatic question generation for educational applications – the state of art. In van Do, T., Thi, H.A.L., Nguyen, N.T. (Eds.) *Advanced computational methods for knowledge engineering* (pp. 325–338). Cham: Springer International Publishing.

Lee, C.H., Chen, T.Y., Chen, L.P., Yang, P.C., Tsai, R.TH. (2018). Automatic question generation from children's stories for companion chatbot. In: 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 491–494. https://doi.org/10.1109/IRI.2018.00078.

Leo, J., Kurdi, G., Matentzoglu, N., Parsia, B., Forege, S., Donato, G., Dowling, W. (2019). Ontology-based generation of medical, multi-term MCQs. International Journal of Artificial Intelligence, in Education. https://doi.org/10.1007/s40593-018-00172-w.

Liang, C., Yang, X., Wham, D., Pursel, B., Passonneau, R., Giles, C.L. (2017). Distractor generation with generative adversarial nets for automatically creating fill-in-the-blank questions. In: the Knowledge Capture Conference, p. 33. https://doi.org/10.1145/3148011.3154463.

Liang, C., Yang, X., Dave, N., Wham, D., Pursel, B., Giles, C.L. (2018). Distractor generation for multiple choice questions using learning to rank. In: the 13th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 284–290. https://doi.org/10.18653/v1/W18-0533.

Lim, C. S., Tang, K. N., Kor, L. K. (2012). Drill and practice in learning (and Beyond), Springer US, pp. 1040–1042. https://doi.org/10.1007/978-1-4419-1428-6_706.

Lin, C., Liu, D., Pang, W., Apeh, E. (2015). Automatically predicting quiz difficulty level using similarity measures. In: the 8th International Conference on Knowledge Capture, ACM.

Lin, C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. In: the Workshop on Text Summarization Branches Out.

Liu, M., & Calvo, R.A. (2012). Using information extraction to generate trigger questions for academic writing support. In: the International Conference on Intelligent Tutoring Systems, Springer, pp. 358–367. https://doi.org/10.1007/978-3-642-30950-2_47.

Liu, M., Calvo, R.A., Aditomo, A., Pizzato, L.A. (2012a). Using Wikipedia and conceptual graph structures to generate questions for academic writing support. *IEEE Transactions on Learning Technologies*, *5*(3), 251–263. https://doi.org/10.1109/TLT.2012.5.

Liu, M., Calvo, R.A., Rus, V. (2012b). G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse*, *3*(2), 101–124. https://doi.org/10.5087/dad.2012.205.

Liu, M., Calvo, R. A., Rus, V. (2014). Automatic generation and ranking of questions for critical review. *Journal of Educational Technology & Society*, *17*(2), 333–346.

Liu, M., Rus, V., Liu, L. (2017). Automatic Chinese factual question generation. *IEEE Transactions on Learning Technologies*, *10*(2), 194–204. https://doi.org/10.1109/TLT.2016.2565477.

Liu, M., Rus, V., Liu, L. (2018). Automatic Chinese multiple choice question generation using mixed similarity strategy. *IEEE Transactions on Learning Technologies*, *11*(2), 193–202. https://doi.org/10.1109/TLT.2017.2679009.

Lopetegui, M.A., Lara, B.A., Yen, P.Y., Çatalyürek, Ü.V., Payne, P.R. (2015). A novel multiple choice question generation strategy: alternative uses for controlled vocabulary thesauri in biomedical-sciences education. In: the AMIA annual symposium, american medical informatics association, pp. 861–869.

Majumder, M., & Saha, S.K. (2015). A system for generating multiple choice questions: With a novel approach for sentence selection. In: the 2nd workshop on natural language processing techniques for educational applications, pp. 64–72.

Marrese-Taylor, E., Nakajima, A., Matsuo, Y., Yuichi, O. (2018). Learning to automatically generate fill-in-the-blank quizzes. In: the 5th workshop on natural language processing techniques for educational applications. https://doi.org/10.18653/v1/W18-3722.

Mazidi, K. (2018). Automatic question generation from passages. In Gelbukh, A. (Ed.) *Computational linguistics and intelligent text processing* (pp. 655-665). Cham: Springer International Publishing.

Mazidi, K., & Nielsen, R.D. (2014). Linguistic considerations in automatic question generation. In: the 52nd annual meeting of the association for computational linguistics, pp. 321–326.

Mazidi, K., & Nielsen, R. D. (2015). Leveraging multiple views of text for automatic question generation. In Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (Eds.) *Artificial intelligence in education* (pp. 257–266). Cham: Springer International Publishing.

Mazidi, K., & Tarau, P. (2016a). Automatic question generation: From NLU to NLG Micarelli, A., Stamper, J., Panourgia K. (Eds.), Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-39583-8_3319-39583-8_3.

Mazidi, K., & Tarau, P. (2016b). Infusing NLU into automatic question generation. In: the 9th International Natural Language Generation conference, pp. 51–60.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*(4), 235–244.

Mitkov, R., & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. In *The HLT-NAACL 03 workshop on building educational applications using natural language processing, association for computational linguistics, pp. 17–22*.

Mitkov, R., Le An, H., Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural language engineering*, *12*(2), 177–194. https://doi.org/10.1017/S1351324906004177.

Montenegro, C. S., Engle, V. G., Acuba, M. G. J., Ferrenal, A. M. A. (2012). Automated question generator for Tagalog informational texts using case markers. In *TENCON 2012-2012 IEEE region 10 conference, IEEE, pp. 1–5*. https://doi.org/10.1109/TENCON.2012.6412273.

Mostow, J., & Chen, W. (2009). Generating instruction automatically for the reading strategy of self-questioning. In: the 14th international conference artificial intelligence in education, pp. 465–472.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobin, B., Valeri, J. (2004). Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology Instruction Cognition and Learning*, *2*, 97–134.

Mostow, J., Yt, H.uang., Jang, H., Weinstein, A., Valeri, J., Gates, D. (2017). Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children's comprehension while reading. *Natural Language Engineering*, *23*(2), 245–294. https://doi.org/10.1017/S1351324916000024.

Niraula, N.B., & Rus, V. (2015). Judging the quality of automatically generated gap-fill question using active learning. In: the 10th workshop on innovative use of NLP for building educational applications, pp. 196–206.

Odilinye, L., Popowich, F., Zhang, E., Nesbit, J., Winne, P.H. (2015). Aligning automatically generated questions to instructor goals and learner behaviour. In: the IEEE 9th international conference on semantic computing (ICS), pp. 216–223. https://doi.org/10.1109/ICOSC.2015.7050809.

Olney, A. M., Pavlik, P. I., Maass, J. K. (2017). Improving reading comprehension with automatically generated cloze item practice. In André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (Eds.) *Artificial intelligence in education* (pp. 262-273). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_22.

Papasalouros, A., & Chatzigiannakou, M. (2018). Semantic web and question generation: An overview of the state of the art. In: The international conference e-learning, pp. 189–192.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. In: the 40th annual meeting on association for computational linguistics, Association for computational linguistics, pp. 311–318.

Park, J., Cho, H., Sg, L. (2018). Automatic generation of multiple-choice fill-in-the-blank question using document embedding. In Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B. (Eds.) *Artificial intelligence in education* (pp. 261–265). Cham: Springer International Publishing.

Patra, R., & Saha, S.K. (2018a). Automatic generation of named entity distractors of multiple choice questions using web information Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J. (Eds.), Springer, Berlin.

Patra, R., & Saha, S.K. (2018b). A hybrid approach for automatic generation of named entity distractors for multiple choice questions. Education and Information Technologies pp. 1–21.

Polozov, O., O'Rourke, E., Smith, A. M., Zettlemoyer, L., Gulwani, S., Popovic, Z. (2015). Personalized mathematical word problem generation. In *The 24th international joint conference on artificial intelligence (IJCAI 2015), pp. 381–388*.

Qayyum, A., & Zawacki-Richter, O. (2018). Distance education in Australia, Europe and the americas, Springer, Berlin.

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In: the 2016 conference on empirical methods in natural language processing, pp. 2383–2392.

Rakangor, S., & Ghodasara, Y. R. (2015). Literature review of automatic question generation systems. *International Journal of Scientific and Research Publications*, *5*(1), 2250–3153.

Reisch, J. S., Tyson, J. E., Mize, S. G. (1989). Aid to the evaluation of therapeutic studies. *Pediatrics*, *84*(5), 815–827.

Rocha, O.R., & Zucker, C.F. (2018). Automatic generation of quizzes from DBpedia according to educational standards. In: the 3rd educational knowledge management workshop (EKM).

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., Moldovan, C. (2012). A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse*, *3*(2), 177–204.

Rush, B. R., Rankin, D. C., White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*(1), 250. https://doi.org/10.1186/s12909-016-0773-3.

Santhanavijayan, A., Balasundaram, S., Narayanan, S. H., Kumar, S. V., Prasad, V. V. (2017). Automatic generation of multiple choice questions for e-assessment. *International Journal of Signal and Imaging Systems Engineering*, *10*(1-2), 54–62.

Sarin, Y., Khurana, M., Natu, M., Thomas, A. G., Singh, T. (1998). Item analysis of published MCQs. *Indian Pediatrics*, *35*, 1103–1104.

Satria, A.Y., & Tokunaga, T. (2017a). Automatic generation of english reference question by utilising nonrestrictive relative clause. In: the 9th international conference on computer supported education, pp. 379–386. https://doi.org/10.5220/0006320203790386.

Satria, A.Y., & Tokunaga, T. (2017b). Evaluation of automatically generated pronoun reference questions. In: the 12th workshop on innovative use of NLP for building educational applications, pp. 76–85.

Serban, I.V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., Bengio, Y. (2016). Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. ACL.

Seyler, D., Yahya, M., Berberich, K. (2017). Knowledge questions from knowledge graphs. In: The ACM SIGIR international conference on theory of information retrieval, pp. 11–18.

Shah, R., Shah, D., Kurup, L. (2017). Automatic question generation for intelligent tutoring systems. In: the 2nd international conference on communication systems, computing and it applications (CSCITA), pp. 127–132. https://doi.org/10.1109/CSCITA.2017.8066538.

Shenoy, V., Aparanji, U., Sripradha, K., Kumar, V. (2016). Generating DFA construction problems automatically. In: The international conference on learning and teaching in computing and engineering (LATICE), pp. 32–37. https://doi.org/10.1109/LaTiCE.2016.8.

Shirude, A., Totala, S., Nikhar, S., Attar, V., Ramanand, J. (2015). Automated question generation tool for structured data. In: International conference on advances in computing, communications and informatics (ICACCI), pp. 1546–1551. https://doi.org/10.1109/ICACCI.2015.7275833.

Singhal, R., & Henz, M. (2014). Automated generation of region based geometric questions.

Singhal, R., Henz, M., Goyal, S. (2015a). A framework for automated generation of questions across formal domains. In: the 17th international conference on artificial intelligence in education, pp. 776–780.

Singhal, R., Henz, M., Goyal, S. (2015b). A framework for automated generation of questions based on first-order logic Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (Eds.), Springer International Publishing, Cham.

Singhal, R., Goyal, R., Henz, M. (2016). User-defined difficulty levels for automated question generation. In: the IEEE 28th international conference on tools with artificial intelligence (ICTAI), pp. 828–835. https://doi.org/10.1109/ICTAI.2016.0129.

Song, L., & Zhao, L. (2016a). Domain-specific question generation from a knowledge base. Tech. rep.

Song, L., & Zhao, L. (2016b). Question generation from a knowledge base with web exploration. Tech. rep.

Soonklang, T., & Muangon, W. (2017). Automatic question generation system for English exercise for secondary students. In: the 25th international conference on computers in education.

Stasaski, K., & Hearst, M.A. (2017). Multiple choice question generation utilizing an ontology. In: the 12th workshop on innovative use of NLP for building educational applications, pp. 303–312.

Susanti, Y., Iida, R., Tokunaga, T. (2015). Automatic generation of English vocabulary tests. In: the 7th international conference on computer supported education, pp. 77–87.

Susanti, Y., Nishikawa, H., Tokunaga, T., Hiroyuki, O. (2016). Item difficulty analysis of English vocabulary questions. In *The 8th international conference on computer supported education (CSEDU 2016), pp. 267–274*.

Susanti, Y., Tokunaga, T., Nishikawa, H., Obari, H. (2017a). Controlling item difficulty for automatic vocabulary question generation. *Research and Practice in Technology Enhanced Learning*, *12*(1), 25. https://doi.org/10.1186/s41039-017-0065-5.

Susanti, Y., Tokunaga, T., Nishikawa, H., Obari, H. (2017b). Evaluation of automatically generated English vocabulary questions. Research and Practice in Technology Enhanced Learning 12(1). https://doi.org/10.1186/s41039-017-0051-y.

Tamura, Y., Takase, Y., Hayashi, Y., Nakano, Y. I. (2015). Generating quizzes for history learning based on Wikipedia articles. In Zaphiris, P., & Ioannou, A. (Eds.) *Learning and collaboration technologies* (pp. 337–346). Cham: Springer International Publishing.

Tarrant, M., Knierim, A., Hayes, S. K., Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, *6*(6), 354–363. https://doi.org/http://dx.doi.org/10.1016/j.nepr.2006.07.002.

Tarrant, M., Ware, J., Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, *9*(1), 40. https://doi.org/10.1186/1472-6920-9-40.

Thalheimer, W. (2003). The learning benefits of questions. Tech. rep., Work Learning Research. http://www.learningadvantage.co.za/pdfs/questionmark/LearningBenefitsOfQuestions.pdf.

Thomas, A., Stopera, T., Frank-Bolton, P., Simha, R. (2019). Stochastic tree-based generation of program-tracing practice questions. In: the 50th ACM technical symposium on computer science education, ACM, pp. 91–97.

Vie, J. J., Popineau, F., Bruillard, É., Bourda, Y. (2017). A review of recent advances in adaptive assessment, Springer, Berlin.

Viera, A. J., Garrett, J. M., et al. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, *37*(5), 360–363.

Vinu, E.V., & Kumar, P.S. (2015a). Improving large-scale assessment tests by ontology based approach. In: the 28th international florida artificial intelligence research society conference, pp. 457–462.

Vinu, E.V., & Kumar, P.S. (2015b). A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. *Web Semantics: Science, Services and Agents on the World Wide Web*, *34*, 40–54. https://doi.org/10.1016/j.websem.2015.05.005.

Vinu, E.V., & Kumar, P.S. (2017a). Automated generation of assessment tests from domain ontologies. *Semantic Web Journal*, *8*(6), 1023–1047. https://doi.org/10.3233/SW-170252.

Vinu, E.V., & Kumar, P.S. (2017b). Difficulty-level modeling of ontology-based factual questions. Semantic Web Journal In press.

Vinu, E. V., Alsubait, T., Kumar, P.S. (2016). Modeling of item-difficulty for ontology-based MCQs. Tech. rep.

Wang, K., & Su, Z. (2016). Dimensionally guided synthesis of mathematical word problems. In: the 25th International Joint Conference on Artificial Intelligence (IJCAI), pp. 2661–2668.

Wang, K., Li, T., Han, J., Lei, Y. (2012). Algorithms for automatic generation of logical questions on mobile devices. *IERI Procedia*, *2*, 258–263. https://doi.org/10.1016/j.ieri.2012.06.085.

Wang, Z., Lan, A.S., Nie, W., Waters, A.E., Grimaldi, P.J., Baraniuk, R.G. (2018). QG-net: a data-driven question generation model for educational content. In: the 5th Annual ACM Conference on Learning at Scale, pp. 15–25.

Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher*, *31*(3), 238–243. https://doi.org/10.1080/01421590802155597.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Tech. rep.: National Institute for Science Education.

Welbl, J., Liu, N.F., Gardner, M. (2017). Crowdsourcing multiple choice science questions. In: the 3rd workshop on noisy user-generated text, pp. 94–106.

Wita, R., Oly, S., Choomok, S., Treeratsakulchai, T., Wita, S. (2018). A semantic graph-based Japanese vocabulary learning game. In Hancke, G., Spaniol, M., Osathanunkul, K., Unankard, S., Klamma, R. (Eds.) *Advances in web-based learning – ICWL*, (Vol. 2018 pp. 140-145). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-96565-9_14.

Yaneva, V., et al. (2018). Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In: the 13th workshop on innovative use of NLP for building educational applications, pp. 389–398.

Yao, X., Bouma, G., Zhang, Y. (2012). Semantics-based question generation and implementation. *Dialogue & Discourse*, *3*(2), 11–42.

Zavala, L., & Mendoza, B. (2018). On the use of semantic-based AIG to automatically generate programming exercises. In: the 49th ACM technical symposium on computer science education, ACM, pp. 14–19.

Zhang, J., & Takuma, J. (2015). A Kanji learning system based on automatic question sentence generation. In: 2015 international conference on asian language processing (IALP), pp. 144–147. https://doi.org/10.1109/IALP.2015.7451552.

Zhang, L. (2015). *Biology question generation from a semantic network*. PhD thesis: Arizona State University.

Zhang, L., & VanLehn, K. (2016). How do machine-generated questions compare to human-generated questions?. Research and Practice in Technology Enhanced Learning, 11(7). https://doi.org/10.1186/s41039-016-0031-7.

Zhang, T., Quan, P., et al. (2018). Domain specific automatic Chinese multiple-type question generation. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 1967–1971*. https://doi.org/10.1109/BIBM.2018.8621162.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Ghader Kurdi[1] · **Jared Leo**[1] · **Bijan Parsia**[1] · **Uli Sattler**[1] · **Salam Al-Emari**[2]

Jared Leo
jared.leo@manchester.ac.uk

Bijan Parsia
bijan.parsia@manchester.ac.uk

Uli Sattler
uli.sattler@manchester.ac.uk

Salam Al-Emari
salam.ammari@uqu.edu.sa

[1] Department of Computer Science, The University of Manchester, Manchester, UK

[2] Umm Al-Qura University, Mecca, Saudi Arabia